Erik: Joining me now is freelancer.com founder Matt Barrie. As usual, Matt has written an excellent paper about AI. This one is called Pay to Pray. That's pay to PRAI. You can find that linked in your research roundup email. If you don't have a research roundup email, it means you're not yet registered@macrovoices.com.

Just go to our homepage, macro voices.com. Look for the red button above Matt's picture on the homepage that says, looking for the downloads, Matt, before we even. Get into all of what's going on in AI. We've got some news just as we're recording this on Tuesday evening US time. There's some recent news just in the last day or so, which is, a lot of companies before they, IPO, they'll do a little bridge round, eh, 10, 20 million bucks.

Cover some expenses until they get to their actual IPO. OpenAI is planning an IPO. They did a little bridge round, 122 billion with a B dollars all time record for a private fundraising round. And what is that? Something like a hundred times bigger than any private fundraising round has ever occurred like before 2025.

Why did they need 122 billion just to bridge them from now to their IPO, which is expected in less than a year.

Matt: Thanks for having me. It's truly is stupendous. But before 2025 the largest sort of, private venture rounds were in the single digit billions

The headline number is 122 billion raised on a 730 billion pre-money valuation. When you get down to the actual segmentation of kind of what's going on, it's it seems that it's only about $25 billion worth of cash.

It's, and it seems to be more of a vendor financing than an actual straight cash injection. You've got Amazon, Nvidia and, SoftBank primarily putting the money in or the in kind in Amazon's putting in 50 billion but that's contingent on OpenAI spending a hundred billion I think over the next eight years on their compute, which I'm sure what Microsoft happy.

And it seems like a good deal, at least on on paper for Amazon. So there's 15 billion going in upfront and 35 Billion Is further more contingent on the company, either going public by 2028 or achieving artificial general intelligence. I'm not sure how they're going to really define that. I think the underlying definition is a panel of experts will make a decision, yes or no which is a bit strange for a financial decision.

There's 15 billion down upfront, 35 billion contingent on going public or AGI for a hundred billion commitment the other way round. So it's a bit like a procurement round. SoftBank loves doing the sort of, I like to call them sort of Russian stents where they lead rounds and mark up valuations to the moon.

We saw it with WeWork. We're seeing it again with Enron, OpenAI. And they've already putting about 40 billion, they're putting another 30 billion in, but to raise that money. They don't actually have the cash on the balance sheet. So they've taken a 12 month bridge loan for 40 billion.

And they're trenching in 10 billion at a time over the course of the, over the year for a total 30 billion. And obviously that's just getting them through. So to the IPOs, they've got liquidity event. And then Nvidia is putting it in kind as they as. Tend to be doing in all these sort of, circular economy, deals in the AI space where, GPUs and infrastructure will be provided to the tune of 30 billion into the round.

So it's about it's actually a $25 billion round of cash upfront. And which is 10 from 10 from SoftBank, 15 from from Amazon. We'll see if the rest of the money comes in, but the rest is in kind. So it seems from looking at this. If you look at the compute numbers, they truly are astronomical that are being contributed in the form of either Nvidia credits, I think it's three gigawatts of inference and two gigawatts of training capacity as part of the, a part of this investment.

That's the power that gets drawn by a small country. So it seems to be what they're trying to do is Scale up the the spend on compute so much that they can find, a way to bring the unit economics down. Because that's really the key problem in the space. Nobody, in the AI compute space is making any money other than really Nvidia. Who does about 160 billion of, of revenue and a hundred billion of earnings.

And then TSMC provides the chips to Nvidia. But the rest of the space is actually negative on using the products in terms of the unit economics. So the more you use the product, the more you lose the money. So I think they're trying to make it up on volume.

Erik: Matt, I can't imagine any responsible business executive signing off on $122 billion deal unless the underlying business model was rock solid.

It didn't have any major risks in it just happens that you wrote dismissive over the last couple of weeks called Pay to Pray. I don't think that's actually the

conclusion that you reached. Tell us about the economics of the, I'll call it the consumer AI business model, offering AI through chat to people like me who sign up for a max subscription on Claude, or a pro subscription on OpenAI.

How much money are they making or losing on that?

Matt: The AI industry is consuming an absolute bonfire of money. It's about $600 billion a year that's being spent by the hyperscalers on CapEx. It's it's reaching a kind of a point now, which is incredible. Where, the CapEx is higher than the kind of internal free cash flow.

The fundamental business model that's being pushed in the consumer market and the software development market. Up until now has really been a venture capital style subsidized model where you pay, you either use the free product and then hopefully upgrade to the 20 a month product, or you use the 20 a month product if you're a consumer or a 200$ a month product, if you're a power user or if you're starting to do programming.

But the problem with with these models are that they take an incredible amount of money to train. And I think we talked about that in a previous episode. Where you got training running North of a hundred million dollars a run, approaching half a billion dollars a training run that requires a huge amount of data center build out a whole incredible stupendous amount of data needs to go into these models.

And because the cheap data that's scraped off the internet for free is basically drilled out to an extent. They have to do licensing deals to get access to that data. Sometimes they do dodgy things like anthropics, scraped a whole bunch of books and scan them in.

And they got caught and they had to pay the biggest fine. I think in copyright history, I think it's about 1.5$ billion for the illegal scanning of all that data, et cetera. So it's incredibly expensive to train. But the fundamental problem is that when you actually use the inference, you basically put queries into GPT or queries into Claude and you run them.

Those queries are loss making. You can't make it up on volume under the current models. While we do see, the underlying hardware is on a Moore's law sort of trend and Moore's Law, for those of you that dunno what it's is, every 18 months to two years or so effectively the technology that goes into chips allows semiconductors to be used with finer and finer featured sizes, which effectively allows the compute capability in terms of processing power to, to effectively

double the cost to have every 18 months while. you are waiting on that silicone Moores law trend because the the models. Are in such a brutally competitive environment where there's literally zero lock in. There's very, almost zero switching costs. So I can, if one day Chat GPT 5 comes out, I'll switch to that, but then Claude 4.6 Opus comes out and that's better.

I'll switch to that. And, so there's nothing stopping me overnight really just changing the models because of, because of the competitive environment. And, and so forth. What the situation basically is that each generation of model is being rushed out to get the top of the scoreboard.

So that all the customers, flock to that. And as a result of that, the actual amount of inference or tokens burned, if you will per useful query with each generation of model. While the Underlying compute is getting cheaper and cheaper in terms of what you can get done on the chips.

The amount of inference you have to burn for a useful query is actually going up quite dramatically. And so you're actually not seeing a reduction in the cost of inference. You're se you're seeing an increase, and that's before you consider the issue with energy and, constrained your build capacity for building data centers and power gear and all the other things.

And a few people have done various models of how much it costs for GPT to deliver their 20$ Plans. And even if you ask Claude itself, you just type a query into Claude saying, how much is the underlying compute costs for you on a 20 plan? With Claude, it'll tell you 15 to $20, maybe 80, 18 maybe a bit more.

So effectively there's no money being made on these $20 plans. And when you're a power user you can burn up to several hundred dollars on the, on the 20$ plan and it gets even worse on these 200$ plans. Which is used by programmers now because, the nature of programing you are a stream of tokens almost forever and the average user on. On a 200$ plan can burn, many thousands of dollars of underlying compute. And in fact, there's a leaderboard called vibe rank, which where people compete to see how much money how much underlying compute they can burn on a 200 plan. The leading guy on the on the leaderboard has burn 51,000 US dollars in a single month on a 200 plan.

So the issue is that as these models get more and more competitive and new versions come out, you're getting an exponential increase in the amount of inference you've gotta burn. And that is, is meaning that these models are not getting cheaper and you're not making up on unit economics.

And so it feels like this funding round is, especially with the amount of vendor financing that's packed into it, it feels like this is an attempt to really try and scale up. The underlying infrastructure and GPU capabilities so that potentially some sort of threshold can be crossed in the underlying economics.

So inference can be profitable, but there's one big problem with all of that which is the demand there and and that's where we are.

Erik: Matt, you said it's $122 billion capital raise on a pre-money of 730. So I get 852 billion is the current enterprise value, assuming an up round, which is what everybody assumes, eh, we're gonna go just, let's call it an even trillion for the IPO 'cause hey, what's, a hundred billion here, a hundred billion there. Eventually you're talking about real money, but holy cow. If, let's say that, that OpenAI. Goes ahead and IPOs sometime this year for $1 trillion. Are you going to be long, short, or flat? And why?

Matt: It really does feel in the space that we're. Really in the moment where a supernova is starting to explode, and we're probably gonna end up with a giant black hole like we did in the dotcom boom the first time around.

That's a trillion dollar valuation at IPO is absolutely gigantic. They've raised 122 billion here in this round. They were mooting that the IPO earlier in the year was gonna be a 60 billion raise in about a trillion. You could probably imagine that number is possibly up a little bit, or they wanna keep it tight, obviously, because once it gets the public markets, you don't have too much stock unload onto the market, but.

A, a bigger problem for them is the fact that you've got Elon Musk in the wings who's not just suing OpenAI because it turns out you can't IPO a charity. And they've converted OpenAI to a for profit model and there's a lot of complications around that. But he's also IPOing his SpaceX for, I think it's 1.75 trillion valuation.

So there's a big possibility that a lot of the heat's gonna be taken out the market when he does his 70 billion, 1.75 trillion raise, and it looks like he's gonna beat OpenAI To, to, to going public because you, if you look at the news and you know what's all happening with the ETFs and the allocations and so forth he's a lot further along.

I, I dunno in the olden days, if you would invest in Amazon when it went public, or Microsoft when it went public, you had the ability to make a lot of money. I, how much is left on the table for the general public when the

company is being valued at a Trillion dollars when it goes public, know, whatcha gonna do get to 2 trillion, 10 trillion, I think Nvidia is what?

Four? 4 trillion. Valuation and they're the only ones making money in the space.

Erik: Matt, the parallels between this and the late 1990's Dotcom boom before the 2000's dotcom bust are just striking to me and it occurs to me before I go on that we probably have listeners that weren't even born when that happened.

So for anyone who's not familiar with what happened there. Wall Street became absolutely obsessed with the idea that the internet is going to be a really big deal. And the thing to I, it's really important to understand about this, is they got that call exactly right. The internet. The public internet was going to change the world we lived in ways beyond what anyone could even conceive.

And they got the call right, that it was a really big deal, and I think they're getting the call right. Again, that AI is a really big deal, maybe as big or bigger than the public internet. But the thing is, even though they got the call right. They started throwing money at dumb ideas without thinking, and it just turned into a complete frenzy.

It seems like that's happening again here. But in the late nineties, it was every tiny little company that had.com in its name, and it didn't really matter whether they had a business model. It's. Different here. It's the big players, which frankly have a business model.

But as you very eloquently explain in this excellent piece, I recommend everyone read called Pay to Pray that business model isn't viable for the reasons that you just described, or it's not profitable or not likely to be sustainable. How should we think about this? I, is it inevitable that a.com bust like we had in 2000 is coming for AI and does it have the same dimensions as the one in 2000?

Does? Is it bigger, smaller, worse, better? What do you think?

Matt: So if you think about the.com boom, and I was there actually in Silicon Valley in 97,98,99 2000. I saw it all. The, there's a, the hypothesis, the internet was gonna be a big thing and it turned out to be a enormous thing in terms of the benefits to humanity and society.

And we'll continue to grow in terms of it as its applications. And AI is the same thing. A AI is gonna be absolutely transformative for humanity in terms of in

terms of what can do, at that time, a company that whose valuation went through the roof was Cisco. And their tagline was, we Network networks.

We talked about, I think at Last macro Voices, every time you plugged in a, a bit of the internet, you needed to have a a router to connect up the network. And Cisco equipment was gonna be everywhere. And why wouldn't it be the most stable company in the world? The same as with OpenAI to an extent, you think, okay, they've got the best AI models, AI's gonna be everywhere.

Why wouldn't AI OpenAI be the most valuable company in the world? But then at the same time, you also had At&t. And up until about 1996 AT&T had a 60% market share it. It basically built. The network, using Cisco equipment to connect up the world. You had three players in the market and I think it was mentioned at the time by an analyst that they had margins that would make drug dealers blush.

But the problem was that in 1996, the telecommunications, act, regulation Act came in and deregulated the market. And then you started having fourth entrance and fifth entrance and so forth and people buying, companies buying, selling capacity to each other, et cetera. And when you started having the fourth entrant come in and competing on cost, the unit economics fell apart.

And if you think about the AI compute space. Amazon had a pretty good up until the AI boom. It's, I think it's its CapEx as a percentage of earnings was down to about 6% at one point. Now, these hyperscalers are spending, the cloud computing companies are spending, 60% of earnings on CapEx, over a hundred percent of earnings on CapEx.

They're sending $600 billion a year. At the moment in terms of a run rate in CapEx, and it's hypothesized that by 2030 it's gonna be 5.2 trillion of CapEx. Now when you just had Amazon, really dominating in terms of a cloud. It was sitting pretty and I think I had about, about a 60% market share.

Now you've got Microsoft who's taken over and they their market leader and and and so forth. And now you've got entrant such as Oracle that's competing on price, and you've got core weave in the neo clouds and the unit economics are starting to look a little bit shaky. and then you had a big bust in the bust where, all these companies have money being thrown at 'em, and you're seeing that right now. Any company that kind of has AI in its, business plan is getting these stupid seed rounds in the hundreds of millions of dollars. I remember when seed rounds were hundreds of thousands of dollars now, hundreds of millions of dollars.

And you had the, through the early two thousands, you had a real trough in the technology space where everything blew up and somewhat uninvestible for a period of time, but. Through that period of time, you had companies like Google and Amazon continue to grow and double down and Microsoft and so forth.

And ultimately today became very big companies. I think we're seeing this on steroids right now with the with AI CapEx and compute and the OpenAI funding rounds and so forth. And it's unquestionable AI is gonna be completely end during, and just absolutely game changing for society.

But I think the kind of circle jerk of money that's sloshing around between a very small number of companies that's stupidly high valuations and it's scale that. It's just stupendous is gonna potentially end up in a big bust.

Erik: Now I want to focus on that because a lot of people are going to mishear what you just said, and they're gonna say, Matt Barrie said imminently tomorrow, there's about to be a great big bust in all of the AI stocks.

That's not what you said. And it really rings home for me because the reason that my partners and I sold our software company. In the summer of 1998 is because we knew it was a bubble, and my plan at the time was to take all of the proceeds and short the nasdaq because I knew it was a bubble.

Fortunately, I got talked out of that, but if I hadn't, I would've lost everything because even though we were right, it basically, the NASDAQ doubled between 98 and 2000 before it crashed exactly like I predicted it was going to crash. What do we do here? It's do you go long? Do you go short?

And I'll say I'll ask the question this way. Open. AI and anthropic are both expected to IPO probably in 2026. When they do, is it time to go long or is it time to go short because I could make either argument. It seems to me eventually you wanna be short, but how do you time that?

Matt: That's the trillion dollar question, right? You've got SpaceX going public, which obviously has GR and XX AI within it. You've got anthropic going public and you've got OpenAI going public. That's why I think it feels like a bit like a supernova. So we'll end up with a big bang in one way, one way or another.

I think the issue's gonna be, some of these valuations, like the OpenAI valuation is predicated on the fact that it's gonna capture. An enormous amount of value out of the world in order to justify these valuations. These funding rounds are so

large and the valuations are so high that you need to pitch a revenue line that that, that matches them.

And, at the moment, open is doing about 2 billion a month in revenue. I think it's about a run rate of about 25 billion. I wish they do stop using the word run rate because I think you've actually gotta recur once before you can call something annual recurring revenue. So it's doing about 2 billion a month of revenue losing 14 billion this year.

I think the burn rate that came out today was about, is expected to lose $70 million a day this year, scaling to 56 million a day, lost. Next year. But in order to to, cross the value of death and get to the promised land, they've gotta, they've gotta show a business model. Makes sense. And I think what their business model is gonna is twofold.

One is that they're going to take. A substantial amount of white collar jobs away from humans. And some substantial percentage of jobs in the world are gonna go to ai and that's the top line revenue number. And then the, in terms of generating earnings out the other side the justification is that the infrastructure that's being contributed as part of this 22 billion round is at such a scale.

That only OpenAI will have the unit economics that will make the inference profitable. So in a, in combination with taking everyone's job and having the only infrastructure that can run this sort of stuff profitably, I think that's the argument for justifying these sort of trillion dollar valuations now.

I don't think OpenAI is gonna capture, for example, hypothetically the value in the AI powered drug discovery market any more than At&t captured the value of the iPhone. The underlying technology is phenomenal. But as Ilya Sutskever said himself he was one of the founders of OpenAI that you can just read 40 academic papers.

95% of what's out there in AI is published in, in, in the public domain. That's why. Every couple of weeks there, there's a kind of a new foundational model that's top of the leaderboard. And every once in a while there's a team of people that you have never heard of 160 engineers in the room, on Guangzhou like, like deepseek, that comes out nowhere and suddenly, leads in the leaderboard because it turns out that, there's no sustainable competitive advantage in the underlying foundational models.

And what you do need though, is you need access to data. But a lot, a lot of that, those data sets are are now out there and public, or in the case of the Chinese,

they probably don't care too much about copyright. So we always have a sustainable competitive advantage and I think there's incredible and tremendous opportunity in AI, but it's probably not gonna be in the, in these models that, that haven't really figured out what the business model is.

Erik: Matt, I think there is a very distinct difference between this scenario and the late nineties to two thousand.com story, and that is this let's say, we don't know whether it's 97, 98 or 99 right now, but we know that March of 2000 is coming. At some point, there's gonna be a washout in this commercial AI space where they just realized that.

The business model wasn't sustainable and it all starts to fall apart. What happened in the.com bust is we went really a good solid couple of years of kind of a technology recession where there wasn't a whole lot of progress on the public internet because we had to basically shake that mal investment out of the system, get back to efficient capital allocation.

And then, after 2002, 2003, things started to really take off again. Okay. What if. The US government steps in and says, no, wait a minute. The military applications of AI create an existential threat to the country. If we don't stay ahead of this, we can't tolerate a 2000 to 2003 pause. You must continue.

You must do it under US government funding, but we're not gonna fund the giving away stuff for free too. Max subscribers on Claude, we're gonna take it over. And, it's just for the military now. We're not gonna have consumer AI anymore. Is that a realistic scenario or do we need consumer AI in order to train the models in order for the military to get the benefit?

What would happen in that scenario where the military says, no we can't allow a stock market crash to slow down progress on the military applications of ai.

Matt: I don't think it's realistic that we're not gonna have consumer ai. The Chinese are open sourcing their models and in fact they've got their strategy to open source both the software and the hardware.

And there's been previous leaks Meta's Llama and, so forth so I think, it's here to stay. I think the interesting thing here is that it won't be the government stepping in. I think that there's one big company that's been sitting in the wings that has God awful ai AI product and it's everyone's pockets.

And that's, apple on the iPhone. I think in, in, in our last chat we talked about how God damn awful Siri is for supposedly being your kind of your original

chat bot, AI assistant. And we hypothesized it'll continue to be awful in the next year. And here we're and it continues to be.

Pretty bad, but bad. They've avoided this whole getting sucked into this whole CapEx bonfire. And they've really got two levers they can pull. One is, one is they're sitting on this cash and they've they can sit back and watch. And as with the dotcom bust, there was a lot of infrastructure that went through to the second and third owners where, maybe the first owner that builds out the optical fiber.

Network goes bust and the second owner comes in and tries, maybe makes it to be break even after, after, pur purchasing it at a cheaper price and washing out the, washing out the underlying sunk costs. And then the third owner comes in and makes some money out of it.

And I think you might have that situation in the ai compute space where if there is a problem with some of these large foundational model companies. There might be a second or third owner and Apple would be ideal for that. At the same time, what they've been doing.

They've been putting in some AI silicon into their products. And a lot of this compute that's currently being done in data centers by OpenAI, by Anthropic et cetera, is gonna go to the edge, and it's gonna go to the edge for a variety of different reasons. One is that you don't want Sam Altman training on your data.

And I, we've talked about before, that I think an emperor has no clothes as was heading into Saas. Where I think large enterprises are start thinking to themselves, you know what? I don't want my data in Google Drive and Gmail, it might get trained on. And you're certainly seeing all the major SaaS companies quietly flicking on a switch in the settings without telling you where they're saying by default.

Now we can train on your data potentially, and then you flick it off, but by then it's too late. Your data's been sucked down. A lot of that compute is gonna go to the edge, it's gonna go to cer, to OnPrem in, in the enterprise on compute devices. So it doesn't go into the cloud.

It's gonna be on your phone, it's gonna be on your MacBook, it's gonna be on your Mac studio or whatever it may be. For privacy, for confidentiality, for latency reasons. And the fact that there's some models now that can, that actually can fit on that silicon. And, while it might be one or two generations away, at some point, I think real soon.

A lot of that load and a lot of that compute that's going to AWS data centers and Azure data centers to run Anthropic and then turn around , to run OpenAI is gonna go to the edge. It's gonna be on Apple's devices, it's gonna be on Apple's, laptops it's gonna be OnPrem. And at the same time, Apple's cached up and might end up being the second or third owner of some of these companies.

Erik: Matt, the title of your missive that you just written, pay to Pray or Pay to Pray. PRAI is actually a reference to the inevitability in your prediction of something called paper token monetization. What does that mean? What, how does that relevant explain what that's about?

Matt: The fundamental business model of Silicon Valley venture capitalists is to try and win markets by financing companies with astronomical amounts of money such that that money gets spent on marketing and a subsidization of the product to such a scale that nobody could compete with these Silicon Valley investing companies or unicorns.

And so total nuclear war is launched on a market. So you can think maybe Uber and free rides in China or whatever it may be, or DoorDash delivering, noodles to people in Indonesia or grab or whatever it may be. So what we have here in the AI space is the subsidization model is that you have a freemium, you have a free product, being GPT or what have you, and then you've got a 20$ a month product. Which I said earlier probably cost 20$ to serve and a 200$ product that probably cost 2000$ a month to serve

If you were you to, this can't continue forever. It's, so the question is gonna be at, can we get the unit economics down to do inference to such a point that these models are profitable before they run out of funding. These companies running outta funding. Now the problem is that as you have as as we've moved into, for example, software development, which consumes a never ending amount of tokens to write code because every company in the world powered by software now, there's this huge token burn that's happening in these $200 plans. And if you were to try and make some sort of reasonable software, like margin, like 80% et cetera, you would've to price these plans, not hundred a month, but maybe thousand or thousand a month, higher.

And the, so at some point. The money is gonna run out and I think they had a near death experience in the last couple of weeks. Anyone who, and I know you're constantly using GPT and Claude, et cetera, and you probably noticed the same thing, in the last couple of weeks, there seems to have been a bit of a panic from these companies where, you log into your $200 plan, you type a couple of queries and then you run outta credit.

And you've always known that these companies aren't making money on the inference because instead of saying put your credit card in and top, top up your credit, it puts you in the naughty corner for seven hours or longer. And so at some point where we the inevitable destination.

For these subscription models which are basically massively subsidized Silicon Valley financed and increasingly debt financed business models is that they're gonna have to move to a per token pricing now, and that is gonna cost a lot of money. I think there was some comments in the last couple of weeks on Reddit where someone who's on a two do $200 plan ran out of credit pretty quickly.

You had to get it done. It had to get something done pretty that night. And so we moved to the API pricing, which is using the programming interface and that was costing 200 an hour instead of hundred a month. And he thought, gee, I've used a couple hours used, 500$, 600$

The problem with that sort of pricing is has several dimensions. The first is that already at 20 a month for a plan, which is 240$ a year, that already prices the product out from Over half world population. The median global income in the world is about two a year. So at two 40 a year, you're already 10% of the pretax income for half the people on the planet, right?

So you, you're already quite expensive. The second point problem here is that, you know these programming models which, realistically to make any sort of margin, need to be priced in the thousands of dollars a month or maybe. Ten Thousand dollars a month. You're gonna, you have a bit of a problem in the fact that these models do hallucinate.

And they, their error when they do hallucinate and throw errors. They're very different from humans. When you use a hire, a freelancer, for example, on my website, you put in some money, okay, I'm gonna pay you to build a website for me. You don't release that milestone until the job is done and.

If the human can't figure out a problem, they'll ask their friends, they'll try and find a more senior engineer for advice. They'll get on our forums and ask other people how to solve problems. They'll browse the internet, they'll he'll clum their way out of solving problems and look, they may take they may ultimately not get there, and you might get frustrated with them and wanna find another developer to do the job.

But ultimately they, their failure modes are very different from AI with ai. Sometimes when it hallucinates, it can do wildly crazy things. About three

weeks ago I had a problem with my VPN on my computer. I was using Claude to help me through figuring out how to get it fixed.

I'm a software developer by background. I've engineer degree from Stanford. I do know how to program quite well. But I veered into PowerShell commands in Windows 11, which I dunno very well, and I was blindly pasting in Claude and made me delete my entire network in stack And, you know, so you have these very crazy failure modes with ai where you either go in loops or it goes away in thinking and it goes and burns, 10 x the inference, try, these reasoning chains, trying to figure out what's going on.

Or it just has these crazy suggestions. And it will look at you in the eye with the, the eyes of a sociopath on a first date in some regards. Trying to, gaslight you into thinking that its answer is is true when it's clearly not the case. It's just hallucinated something rather.

And the issue here is that in a paper token pricing model, it really turns software development into a slot machine. In that, if the if ultimately, you're riding your app and you need to get called to do something, rather you're really pulling the slot machine handle, you dunno how much gonna cost, how much it's gonna cost you by the time the tokens are all burned.

And you dunno if you're actually gonna get a solution at the end of the day. And I guess only in Silicon Valley because they turn software development into, to generate gambling because that's where you end up. And the frustrating thing that I think will happen is when you're on the 200$ plan.

You've got a certain amount of capacity and maybe it tells you to time out or what have you. You kind of know, you are capped at 200$ right? It's frustrating. Runs outta credit, you are going to find alternative ways of doing things. You kind of know what you are up for? When pulling the handle onto the paper token model, it could be, $50 per spin. You may go a circle, you circle 10 times.

There's examples of people over of radical crazy things that claude does to your codebase. And I think people are going to get very very frustrated

If they have to put a coin in the machine, pull a handle every time, and they get a non-deterministic outcome of whether, of whether they're moving forward in a hill climbing sense to this, to their final solution and their final app or the final bit of software being developed or whether they're going in circles, round and round again.

And I think people are gonna get frustrated. They're gonna go try and find open source models. They're gonna try and find alternative ways to get things done. And I think it that's really the problem. While the hallucinations are getting are reducing. As part of, each new generation in terms of the engineering of the infrastructure around the foundational models to reduce those hallucinations.

What is actually happening is the token spend is going up exponentially because you're doing more and more complex things. So you have a much bigger surface area in which you could generate an, so what I mean by that is, while the probability of a failure pulling. Pulling the handle on the slot machine is getting reduced with the, the engineering from each, subsequent generation of model.

The number of handle pools you need to make is going up exponentially. And so you've got a multiplicative effect in terms of in terms of potentially the impact of errors. I just think it's going to the question that basically needs to be solved now with this $122 billion fundraise is can you get the cost to compute way down to get this whole model profitable and will the market tolerate.

Software development is a slot machine.

Erik: A lot has been said already by lots of people about the incredible rate of progress and how quickly AI itself is getting smarter. What I don't think has been discussed enough, and I'd like to get your comment on, is the rate to which. Professionals are becoming dependent on it.

I think more than you know this, it's more addictive than cocaine. I remember our first interview on AI when chat GPT had just come out, and I remember thinking to myself, boy, Matt's really into this stuff. To me it's a novelty, but I don't think I'd ever actually pay 20 bucks a month for it.

It's just. It passed a touring test. Big deal. But I I don't think I'd ever buy a subscription. I'll tell you, Matt, in the last several weeks, there's been a lot of stress in my life because of this war and family that are affected by the war. And waking up double digits down on a percentage basis, on my net worth because of something that happened in the market overnight.

Okay, look, I'm a big boy. I've been through that stuff before. I've been through the 2008 crisis. It's not that big of a deal, but Claude 4.6 was going offline and the server wasn't available, and I was freaking the F out. I couldn't handle it. I was losing, and don't you dare insult me by suggesting that I go back to Chat GPT 5.4. I don't drink Pap blue ribbon and I don't do chat GPT. Okay. I've

gotten to the point where I can't live without Claude. This is, this seems like a risk.

Matt: When Claude went down, it's quite interesting actually because, the first time it had a bit of an outage was when the data centers got blown up in, in Dubai and you have to wonder yourself what potentially was being run in those data centers in the Middle East that caused the outage of Claude.

And I do think those companies did have a bit of a near death experience in the last couple of weeks. In that all of a sudden you had Sam Altman kill SORA, which was this hyped up video modality model where you could generate clips or, the whole point was you're supposed to be able to type in a prompt and get a movie out the other side.

In fact, Disney paid a billion dollars to OpenAI for use of the technology, and I only found out half an hour before a meeting that the whole thing was gonna be canceled. At the same time, Sam Aman killed instant checkout and he killed the he was working on some sort of erotic model as well, which is a bit strange, but I think it's probably one of the big markets is his pornography.

And he thought maybe he could make some money there. He killed that as well. And the same time Claude had this big changes in terms of how they did the plans and they were giving out, extra tokens in off peak, but in peak they're you back, et cetera. And ultimately it seems that they've cut back a token budget models.

On the path to this sort of pay per, pray Pay per token business model. So I do think they all had a bit of a near death experience and that near death experience obviously is what I was feeling. What you've been feeling when you see these models, you start using 'em going, gee, is my access to gonna start getting restricted?

We'll have to pay a lot more money. We'll have to add a zero to, to, to my monthly subscription. What? I'm have to pay per query. What's going on? And then of course you've got this financing round which is less cash and more infrastructure.

And perhaps it's a way to help the company towards IPO. So there's liquidity event. So the, the original investors can make a bit of a return. It turns out that, this whole it's pretty funny that this whole AI compute space is predicated on $600 billion a year of CapEx which is incredibly energy intensive.

When at the same time your bombing the, you know an area Where 48% of the world's energy is. And and you're relying on energy being cheap in order to have this AI boom work.

Erik: Let's move on to private credit. There's a lot that's been said as a lot of private credit funds are gating investors that this is AI driven or it's related to Claude Codes, specifically creating fears that software companies would no longer be profitable.

And I have a hard time with this one because this sounds like the claim I remember hearing in the 1970s when supposedly the introduction of the, hewlett Packard Digital Computer, or edit, please. The introduction of the Hewlett Packard electronic calculator was supposedly gonna put every accountant out of business and create vast unemployment of bookkeepers and so on and so forth, and it was the exact opposite.

It created more productivity. It seems to me that Claude Code just gave the software industry the biggest productivity boosting tool that they've ever had, and that's the reason that private credit is blowing up. Is that really right? Am I missing something?

Matt: There's a few things going on.

First of all, the, these funding rounds are getting too big for equity. As we saw with the hundred 22 billion in all the infrastructure, you the vendor financing in there, and so you've increasingly, these companies are turning to debt. Even Meta had to go and get 30 billion from Blue au not so long ago in order to fund data centers because they can't do it off the balance sheet anymore.

And the numbers are starting to get too big for equity raising. So they're borrowing the money and not just are we seeing record equity rounds. We're seeing record debt rounds that blew our financing of meta was the biggest private credit round ever. the problem is that, private credit in these portfolios has got SaaS businesses, and one of the big pictures that these our compute companies has is that SaaS is dead. Which is ironic when, chat chip on a 20 a month subscription is a SaaS business. So it's funny that they're saying SAS is dead when they actually are SA business themselves but.

There's been some warnings that have come out around these private credit portfolios that pack in the AI debt as well as the SaaS debt because AI is actively pitching. The, it says the future of these SaaS companies will be eroded

by the fact that AI is coming, so it's causing some instability in the private debt markets.

And of course, what else is causing instability is rising interest rates in an uncertain world and a war in Iran amongst other things.

Erik: Matt, speaking of data centers blowing up, let's talk about the Iran conflict, its connection to ai, and particularly in your latest missive. You expressed some concerns that there's a risk that potentially this Iran conflict kind of pulls the rug on the whole AI business model.

What do you mean what's going on?

Matt: This is where the fifth Industrial Revolution meets the Islamic Revolution, right? A lot of the financing for AI has come from the Middle East. And you can imagine now if you're Saudi Arabia or you are the UAE and you have a fiduciary duty to protect your nation and your citizens and your economy, are you gonna be putting it into a hyper rounds of Sam Altman's?

Highly invited, highly inflated highly inflated valuations or will you spend on defense energy rebuilding, rebuilding your civilian and industrial infrastructure and potentially gonna war to to fund an army, right? It's quite problematic when, you know you've got a country that sits right in the middle of, 48% of the world's energy infrastructure controls a strai where 21 million barrels of oil goes through every day that can fire, $20,000 drones at scale into anything within a 2000 kilometers range and

Migrate not just your data centers in the region, but potentially your energy infrastructure into the cloud literally in a puff of smoke.

Erik: Matt, another theme that you have in the latest missive is that you've gotta be pretty smart to really get the most out of ai. What do you mean by that and what are the consequences?

Matt: When I look at deep down at what, what's happening in the space? Despite, you've got a conflation, a few things happening right now. You've got the AI companies trying to justify these huge valuation rounds, and they're doing that by saying they're gonna take away a lot of the world's work.

Then at the same time you've got mass layoffs happening in the market. With Block today with Oracle, I dunno if you saw overnight, but Oracle has announced 18% of their workforce has been cut, Etc And so you would not be surprised that the general market has conflated these things and thinking that AI is taking people's jobs away.

Now. I see AI instead as a tool. It's a very powerful tool. It's yeah, but it's a productivity tool. Much like the world was, when you went to work and there was no computer on your desk and then you went to work and there was a computer at your desk, or you didn't have mobile phones and you have mobile phones, or you didn't have the internet and then you have the internet.

Yes, there is. Incredibly disruptive and transformative time and some jobs are lost, but, just like pretty much all forms of technology, more jobs are created over time. And what I mean essentially by you could be smart to use AI is AI is a power tool and, so as a chainsaw, right? You give a chainsaw to a carpenter and they can do a skilled carpenter, and they do amazing things. You give a chainsaw to a novice and you can cause all sorts of problems, right? And so what I'm observing both across my platform as well as within my my engineering team is the.

The people who are benefiting most from AI are the ones that are highly skilled and intelligent and know how to use the ai, and they're seeing productivity gains, which are astronomical. They're seeing, double to triple their productivity. And you can measure that in different ways in, in terms of what, what they're achieving.

The and then as you go down the skill level you do, it is a rising tide lifts or boats. So if you are an average copywriter, you can now be a good copywriter. If you're an average illustrator, you can be a good illustrator using these various tools. If you, I make a joke, if you're an average programmer, you certainly are a confident programmer now using these tools.

But to really get the best out of them. The people who are highly skilled are the ones that are really driving the outcomes. And I see that, for example with your work and what you do writing these missives and books and so forth. And so I think just just like technology has over time, created a bifurcation in society where you have the people who are skilled and can create technology and that's where all the wealth flows. And that's why you see these companies, huge valuations and and the ultra, high net worth in the technology industry.

And then you see a general deterioration at the low end of society. I think this is going to drive it even further in extent. The people who really know to use AI are gonna capture incredible opportunities in all sorts of different market segments where they can control that customer interface.

Erik: I definitely agree with you.

I just finished a writing project, or I should say Claude, and I just finished a writing project that was considerably bigger than writing my book beyond blockchain. It required a lot more research 'cause beyond blockchain was just my own opinions. It was a lot to it and. Beyond blockchain took me more than three months.

This project took me less than two weeks, and it's just amazing how much you can accomplish, although you definitely, as you say, it takes some skill to learn how to manage context, window exhaustion and recognize symptoms of it occurring and so forth. So I couldn't agree with you more. It scares me though, Matt, because I think.

One of the biggest problems that society faces in the mid 2020s is the K-Shaped economy. The tendency that the rich get richer and the poor get poorer. It sounds to me like AI really is going to exacerbate that in the sense that the smartest people are going to really benefit in productivity from ai.

They're going to be a whole lot smarter and more capable than they used to be. One guy will be able to do the job of 10 or 12 guys, but. The 10 or 12 guys that weren't that bright and really didn't figure out how to recognize the symptoms of context window exhaustion and their LLM I think it really does take their jobs and it seems to me like this could become the basis for deeper division in society.

So Matt, I want to ask you this because as the CEO of freelancer.com, and for any listeners who aren't familiar with Freelancer, it's basically a marketplace where people can hire independent workers, whether they be high-end expert consultants in their field, who are the leaders of their field, or if it's just a guy who will design a logo for you for five bucks, you can hire all those different people on freelancer.

Matt has a very unique perspective on AI because first of all. He's running a company that uses ai. So he leads an engineering team that's building AI based solutions. But some of the people that his customers are hiring are AI experts that are helping companies to implement AI at a corporate level.

And then some of the more mass freelancer, larger group of graphic designers and people that are making logos and so forth are users of AI that are. Leveraging their logo, design contests and all that. So Matt sees all of these different dimensions of different people using AI in different ways, corporations and organizations adopting it.

Versus an individual guy in Indonesia who's just trying to make a buck as a graphic designer suddenly becoming much more productive and being able to work in different. L languages that he doesn't even speak. So Matt sees all of these different dimensions. Matt, from that vantage point that you have, what would you say are the top three things that you've become aware of that the average person who can't see all that stuff probably wouldn't think of.

Matt: The amazing thing is we're gonna see the ability for you to get things done that you could never possibly think of before, right? Rather than just getting a website built, you get a whole business built. The ability for, I think we're gonna enter a whole new world of explosive entrepreneurship where now, you really just need to have an idea to start a company and the ability for you to execute on that using both AI and also accessing humans powered by AI will be unprecedented. You'll be able to do it at cost that's cheaper than ever before. So we're gonna, I think, enter an explosion period of hyper hyper competition. And, thanks to the internet and the ability to distribute products or services, though the internet at scale so quickly, if you've got a great idea, that great idea can take off and you can make a billion dollars faster than any time ever in the history before.

So I think it, it is an incredible time. I certainly haven't seen as yet the complete replacement of someone in a job. I dunno, if you ask around, does anyone know any graphic designers that completely lost their job? Where the dislocation will occur is where you've got highly paralyzable workflows where you've got thousands of people or hundreds of people doing the same job.

So maybe in a call center where there might be 10,000 people or a thousand people in a call center doing the same workflow, which is, customer support, answering our phones, et cetera. With ai, you'll be able to take it from a thousand people down to maybe a hundred people. If you've got 30 junior lawyers drafting legal agreements in a room, maybe you'll be able to take it down to 13 people in a room.

But I don't see because of the way AI works and goes in circles and the failure mode and the fact that as you burn more tokens, you've got more chance, even though the unit rate of errors goes down, the fact you're burning all these. Extra

tokens means that ultimately an error becomes more catastrophic and in a way that a human never would make an error.

What that means is that I think the ultimate combination is humans and ai. I think it's probably one of the productivity to tools known to man, and I think it's gonna open up a whole amazing golden age of building and creating businesses and hyper competition.

Erik: Matt, I can't thank you enough for a terrific interview.

Before I let you go, you run freelancer.com, a public company that trades under ticker symbol FLN on the Australian Stock Exchange. Tell people who are not familiar with it, what Freelancer does, and how to follow your work. Your latest piece, again, it's linked in the research roundup email. It's published on Medium.

It's called Pay to Pray. You write quite a bit of interesting stuff for people who wanna follow your work. And learn more about freelancer.com. Tell us about it.

Matt: We run the world's largest cloud workforce, so there's about 87 million people in that marketplace. We can do any job you can possibly think of from $10 jobs to $10 million jobs.

The mainstream jobs are things like, build me a website or build me a, an app, or it basically help me start a company. Or grow my company. The biggest things we run. We've got a Moonshot Innovation Challenge program at the high end where we help all sorts of US government departments and enterprises around the world solve scientific and technical challenges.

Give us your hardest scientific technical challenge and we'll solve it. Or you don't have, pay the prize out. So the biggest thing we've got running right now is a $10 million. It's a seven half million US dollar gene editing challenge. For in the central n system of humans, which we're doing for the National Institute of Health and and also working with nasa.

Actually, ironically, the when art goes up, in the next 24 hours into space we worked with NASA to crowdsource the mascot from kids all around the world. That's going up with the astronauts to basically inspire kids about about space and so forth. If you wanna get anything done, you come to our site, we can get it done.

And no job is too small, too big, or too complex, and it ranges, through to mechanical engineering or electronic design or whatever you need. And if you wanna follow my writing on ai there's a series of, obviously a podcast that we've done together on Macro voices on ai. I think this may be the fifth or the sixth.

And if you've got a medium or substack, you better find me and see my essays.

Erik: And if you just put Matt's name in Matt Barry at the search box at macrovoices.com, you'll see a list of all of the previous AI interviews that we've done Right here on Macro Voices, Patrick Ceresna and I will Be Back as Macro Voices continues right here at macrovoices.com.