

SM290 | 7.4.2026

How to Raise Your Agent | Episode 8

Matthew Cheung, CEO, ipushpull

This week, we close out our *How to Raise Your Agent* series with Matthew Cheung, CEO of ipushpull. David Greely sits down with Matthew to discuss how chat won the interface war in financial services, how AI is unlocking a trove of chat data for enterprise use, and why governance will be the key to who wins in a world of messaging and AI agents.

[IOSCO paper mentioned during the conversation: “Supervisory Toolkit for AI Use in Capital Markets”](#)

Matthew Cheung (00s):

It's really exciting to see how some of these markets will ultimately start to get electrified. And then I think the piece that is very interesting as agents start to get used more, and this again is some of the problems that are being solved around the identity of those agents and how the permissioning of that works, what they can connect to. And ultimately, if you go forward to, I keep saying that no one is really using those for trading workflows at the moment, they will be probably sooner than we think. But by the end of the decade, no doubt there is going to be agents trading with agents. And how does that work? And I think that's going to be exciting and a little bit scary as well, if you haven't got all of those governance layers.

Announcer (46s):

Welcome to SmarterMarkets, a weekly podcast featuring the icons and entrepreneurs of technology, commodities, and finance ranting on the inadequacies of our systems and riffing on ideas for how to solve them. Together we examine the questions: are we facing a crisis of information or a crisis of trust, and will building Smarter Markets be the antidote?

This episode is brought to you in part by Abaxx Technologies. AI agents are fundamentally changing how people and organizations interact with computers and the internet. A race is underway to establish the path that will govern the nature of these interactions. Abaxx Technologies introduces Abaxx Labs as its center for engaging with the developer community through the release of open source software to promote a path that empowers users to maintain control of their identity, data, and agents in digital spaces. Abaxx Labs' first release is the Agents++ library, a subset of Abaxx's ID++ software development kit that has been tuned for use with AI agents. Download Agents++ on GitHub today at github.com/abaxxlabs. That's github.com/abaxxlabs.

David Greely (02m 06s):

Welcome back to How to Raise Your Agent on SmarterMarkets. I am Dave Greely, Chief Economist at Abaxx Technologies. Our guest today is Matthew Cheung, CEO at ipushpull. We will be discussing how chat won the interface war in financial services, how AI is unlocking a trove of chat data for enterprise use, and why governance will be the key to who wins in a world of messaging and AI agents. Hello, Matt. Welcome to SmarterMarkets.

Matthew Cheung (02m 35s):

Thanks for having me, David.

David Greely (02m 37s):

Well, thanks for joining us. You have spent two decades building communications infrastructure and technology for the financial markets, and I have been looking forward to discussing your thoughts on how AI agents will transform the use of technology and data in these markets. You have been writing on these topics a lot recently. I have seen you have written that AI agents will replace single-purpose applications as the main way people in financial services interact with data. Can you explain a little bit what you mean by that?

Matthew Cheung (03m 14s):

Yeah, sure. I guess it comes down to different interfaces in financial services and markets. People have had GUIs around since the 90s, so people can point and click on the screen. That's evolved into APIs, where you can have systems doing that, and then agents is kind of then the next leap forward. I think that interface theme you have always seen in many different areas of markets. Even in the broader technology space, you had browsers in the 90s, mobile apps in the noughties, and then ChatGPT was the interface for AI. AI has been

around since World War II, but in terms of the Transformer paper, which has then triggered this gen AI boom, that came out five years before ChatGPT came out. It didn't actually come out of nowhere, but then that's what you see with exponential technology. It looks like it's come out from nowhere, but it's slowly getting better over time. So, I think building upon very much that shoulders of giants, which you see with technology, where we are entering this new phase of agents. That's been on the tip of everyone's tongue for the last year or so. Agents are, again, this interface moment into using some of these technologies with the likes of quad code and Anthropic in the enterprise level. It's been booming at the moment. It's been quite interesting. Actually, I was recently at a conference, and even six months ago, when you talk to people in big enterprise, they would talk about, oh, I have just turned on Copilot. We all know Copilot sucks. That's why I would always tell people, if you are in an enterprise, you don't really have much options. I think Copilot is better now, six months later. What's been quite interesting as a trend is that now companies and big enterprises are actually turning on Anthropic and quad code much, much quicker than when they are originally turning on Copilot because there is this race that's on now. I think as time goes on, and we're still in the very, very, very early innings of this, it's still like the 1% or the 0.1% that's really embracing these tools. If anything, it's people that are not in enterprise that are really embracing it because they have got the agility and flexibility to use things like that. I have been quite surprised, actually, with some of the big companies, big banks, big financial institutions that actually have got their hands dirty with quad code. Then you have things talking about how much money should people be spending on tokens and what they're doing and so on. Going back to your question in the medium term, then yes, the ability for someone to orchestrate workloads very easily by interfacing with an agent and their sub-agents and their sub-agents is something that is the direction of travel. There is some probably workflows in financial markets where you still need some element of seeing something. Then that's where I think the level of where the human gets involved in these workflows then changes what the interface or the application or the workflow looks like. It's a really exciting time.

David Greely (06m 27s):

Is the end state of that direction of travel, are we trying to move to an Ironman future where we all become Tony Stark interacting with Jarvis and then Jarvis interacts with everything for us? Is that what it's going to be like instead of having to learn how to code, the agent learns how to speak human?

Matthew Cheung (06m 43s):

I think so. You have got, there is obviously that Ironman analogy from Andre Capote, who also coined the term vibe coding. You have got Ironman and Ironman suits. You have got an Ironman suit with a human inside it as that idea of the human is working collaboratively with AI. You could just have an Ironman suit because in some of those films, the Ironman suit just goes off and does stuff sometimes as well. That's then fully autonomous. I think there is still room for, again, depending on what workflow there is, different levels of input that's required from people, but then human judgment still surpasses what an AI can do in isolation. That's where it's quite interesting if AI starts becoming team members. Do they have their own identity? Do they have their own permissions to do things? Are they just a member of your team? Can they manage other teams? Are you part of their team? All of these things are going to start to change in the next couple of years.

David Greely (07m 41s):

I would love to dig into that with you. I have read that you have written, that you think of the next enterprise era in terms of four pillars, and those being chat, data, AI, and governance. You argue that chat or messaging, which you have spent a lot of time working in, most recently with Ipushpull, that messaging is the interface that won. Why is that and why is it important that messaging is the interface that won?

Matthew Cheung (08m 09s):

Chat and messaging has been around for 50-odd years, since the mid-70s. It really took off in the 90s. You had MSN, and AIM, and Yahoo, and ICQ, and all those type of platforms where people started to use it. That was a lot of fun in those times as well, how those things worked. But then, over time, you have then built again into what you see today. What I find very interesting is, apart from actually using your vocal cords to communicate with another human being, chat is actually the next method that you will use because there is 8 billion people on the planet, and half of them are using chat platforms, or more than half of them are using chat platforms in different shapes and forms. Actually, everyone is used to the big ones like WhatsApp and Microsoft Teams, if you are in a big company but there is actually hundreds of chat platforms out there but all of the usage gravitates towards the big guys. WhatsApp's got 3 billion users, I think now, which has increased rapidly, actually, in the last maybe 24 months or so and even, actually, WhatsApp, if you saw the news earlier this week, I think, they have now got a new CEO of WhatsApp that's actually a former CEO of a very successful Fintech company in India and in India, there is a billion and a half people in India who, not all of them on WhatsApp, because they are probably babies or something, but a good chunk of them are using WhatsApp and if you think about the payment ecosystem in India and you

can integrate that into the most used interface in that country, in a similar way you have seen other stuff happening in China with WeChat.

So I think chat and messaging is integral into enterprise workflows through to talking to your grandparents, right? Everyone is using it, every shape and form but what's quite interesting is all of the chat platforms have been designed for humans to communicate with humans and then what's quite interesting over the last few years is some of the automation that you could start doing within some of these chat platforms and then if you are automating financial services or commodities workflows, then you need obviously all the controls and permissioning around that as well. So going back then to the idea of these four pillars. So from where we sit, so ipushpull very much focuses on chat-driven workflows and the ability to automate them, the ability to extract the value from the untapped kind of asset that is chat data that no one is really using yet. But then because of the convergence of the different technologies we have seen, particularly large language models, Gen AI and so on, it means now having, well Venezuela's got the biggest oil reserve in the world, but you can't tap into it because it's just too expensive. So you don't bother, no one does. And you had a similar thing historically with chat. You think about how many messages are being sent every single second of every single day, but you couldn't do anything with it. So now you've got the tools and the means to do it with Gen AI and large language models and platforms like ipushpull and what we are doing. So then how we kind of see that world is this idea of having these four pillars. So there is chat, data, AI, and governance.

So chat, we have just spoken about, data is then key to everything. So unless you actually have good data, you can't do much, which again, going to that kind of, I suppose, the oil refinery type of analogy, you have got this kind of unrefined, untapped oil that you can't do anything with. You need to extract it out, you need to clean it, you need to refine it, you need to turn it into all these different things that could be usable for other things to do and in a similar way, you have that with chat. So if you imagine in a conversation on a trading desk on any given time, if I am a trader, I might be asking for a price from my dealers and brokers. I might send that out to three of them. I might send them out to 10 different dealers and brokers. And as they are responding in chat, where people are typing with their fingers, I, with my eyeballs, have to watch what's coming back. I will probably have a scratch pad and I am writing them down on a bit of paper and then the reason you're doing that is because that data is completely unstructured and now with data, that kind of pillar that we are looking at, you can take this unstructured chat data, use different techniques to then transform that data into the right message type, message format, enrich it, turn it into different types of messages that could be plugging into different places. So data is very, very important. So that's chat, that's data. AI, already spoken about how that's really, really important with using it as a new tool to extract value out of all of these kind of data and workflows and the important piece is having all of the right then permissioning, access control, identity management, audit, logging, and basically governance.

So governance then is the last pillar, because you can't just be using AI for anything and that's another reason, obviously, why big enterprise, massive, huge tier one banks have been, some of them slower to roll out things just because there is so much going on when you're feeding things into essentially a black box, when you are using some of these frontier models and then that's why it's been, the race is on to find the best way to have all the right controls around it. In Europe, you have seen all the EU AI act has come out to talk about how you need to be record keeping and how you need to have auditability and accountability and all these different processes, laws, governance, everything from the very top level on a regulation level, all the way through to actually what's that lady on the desk who is quoting prices right now, what controls is there on what her agent is doing, right? So it's all of those different layers. There is many, many, many, many layers and governance kind of encapsulates all of them and then where we sit is actually on the application and the execution layer, because that's where kind of the action happens. So yeah, in terms of the workflow, and also we are actually extracting out the Venezuelan kind of oil reserve of chat data. We are part of that and doing the structuring of the data as well. So you need to have that audit trail of every change of what that data has done through different stages, as well as what the workflow has done, who's interacted with it, and so on.

David Greely (14m 36s):

Well, it's such a fascinating way of thinking about it, because I don't know how many folks have thought about the fact that LLMs allow you to unlock this data resource that previously was just too difficult, right? It was too unstructured for people to really be able to sink their teeth into it and make use of it. I guess that creates challenges and opportunities, right? I think a lot of people thought, from a privacy standpoint, people thought, well, chat's kind of ephemeral. Nobody can go through it. Now they can. So there is the historical data that's locked up in texts and WhatsApp and everywhere else and then there is the forward going basis. What do you think about this kind of reservoir of chat data that's been built up over the years? How concerned do people need to be about that now that it can be tapped?

Matthew Cheung (15m 28s):

I focus on the enterprise side. So I will probably put the consumer level side of it away but in terms of the enterprise, in regulated markets, all chat and communications is all monitored. So anytime anyone is having a conversation on the phone, on chat, on email, whatever it is, that is all stored in. There is a big boom in compliance technology, which is archiving and surveillance technology, where the compliance world lives on T +1 . It's not real time. I think they want to get towards real time. But at the moment, and for a long time, it's been T +1. So one day after, they will grab all of this different channels of communication, grab it together. But they don't actually do much with it. They will just store it. And then when the SEC or the FCA or someone gives them a knock on the door one day, they will have to go through the tapes like I used to do in the old days. In my very first job, I had to go through the tapes for someone when we are having to check a trade and I had to go and sit in a server room with these old tapes and literally, we were listening through them for a whole weekend, trying to find out. We are manually transcribing them and so on. So archiving surveillance does a very similar thing where they can then pull that together. So I think in the historical kind of look backwards, now all of a sudden, you can access that. I don't think it's going to make any difference in the immediate term, because they are already storing all this. What will make a difference is the searchability of all of the information, because then rather than human beings having to go through all this stuff, you can then go through stuff. So that's one thing. I think kind of going forward, people don't really need to change what they are doing. A human's workflow should be the same and if you are working in a regulated entity, you know what you should and should not be talking about on your Bloomberg chat, or if you should or should not be using WhatsApp at your desk if you are sitting on a trading floor. And that's all come from regulation. Regulation has been the stick to beat that. So I think the value of what's in that chat is quite interesting, because how when we start with a customer, part of the onboarding that we do with them is actually there is a process called chat data mining, which we use, which then a user of our service will then go through and feed in the unstructured conversations. We then extract out and turn it into structured data and that process of as we are doing it means that you are creating essentially like a library of syntax for that particular trader or broker or desk or organization and then that kind of data model that you are using can then be used as a tool so you could search stuff historically. Otherwise, if you just put a large language model into search for very, very large sets of data, it's not the best way to do it. A large language model has been trained on the text of the internet. It's not been trained on loads of data. So what a large language model likes then is for that data to be standardized and then it's easier for it to access. So I think all of these things are all happening right now, literally in the last year or so. The convergence of these waves are all kind of crashing against each other, which means there is lots of opportunity, lots of people looking at these as hot topics.

David Greely (18m 53s):

And I want to come back to the governance point because you have said that as we look at how this progresses, governance determines who wins and now we have to think about governance not only in terms of humans, but in terms of AI agents and my understanding is you think about this in terms of an autonomy slider. Can you describe the autonomy slider and its role in governance?

Matthew Cheung (19m 17s):

I mentioned Andrej Karpathy earlier. So he was one of the founding members of OpenAI. He was the lead AI person in Tesla, and now he just joined Anthropic actually maybe a month ago or so. A big AI guy, right? So he was talking in a presentation he gave six months ago or so about this idea of an autonomy slider. And you already start to see it. He was using examples of if you use Perplexity, which most people probably do, or if you are using things like cursor if you are coding, there's already a UI where you can kind of change the level of automation that AI tool is doing for you. Now, that idea of this slider in autonomy is not a new concept. It's been around for decades. But then what's happened is kind of evolved over time. Originally it was looking at aircraft cockpits, nuclear power stations, big dangerous stuff back in decades past, where there is a lot of governance and control. If you think about flying on an airplane, it's the safest mode of travel, safer than walking, but that's because of all the governance and controls and auditability and accountability and all of that has happened to it. So that's what needs to be brought into AI and again, even on a political level, you've seen Europe go one way, US has gone another way, OpenAI has gone one way, and Probit's gone another way. So governance is big on that level, but down to an actual workflow level where we look at kind of autonomy in this idea of a slider. So you had the end of May 2026, the International Organization of Securities Commissions, which basically advises global regulators. They released a paper that's worth looking at. You should probably drop in your show notes. It's called Supervisory Toolkit for AI and Capital Markets and then building on some of those terms and concepts that's happened over the years, they then call out, which is fairly standard in the industry, but probably quite new, a broad industry, but quite new for markets. This idea of then those kind of points on the slider is essentially humans in control. So the human's kind of doing everything. You have got a human in the loop, where the AI is acting as like an assistant, but the human still has control of the workflow, and it's just helping the human. You have got a human out of the loop where the AI is doing all of the work, but the human acts as an overseer or an exceptions manager and then you have fully autonomous agents and running off doing lots of things. So in the world of markets, no one is really using autonomous agents when it comes to like trading and things where money's at stake. It will be coming pretty soon. I am sure there is lots of hedge funds that are playing around with it and things like that. But in a

financial institution, the big regulated one, the only level where you will be seeing autonomous agents in the short term will be more around probably post-trade processes, because the trade's already been made. The risk of something is significantly less. So where you sit in the middle then is then at the moment, it's very much the human in the loop and the human out of the loop and currently where the technology is, is sliding probably between those two things.

And then that will be, you may up and down that slider, depending on the complexity of the task, maybe the amount of money that you might be trading with it and four eyes checks and what have you, the level of accuracy, because no large language model gives you 100% accuracy on anything because it's probabilistic by nature, not deterministic. So when it comes to some of the workflows that someone on a trading desk might be doing, you will be looking for, you call it like the three nines, 0.999% accuracy and depending on how much you're using these technologies, you can get it very near to that, but you still then have a human watching and overseeing these things. So that's where, again, I mentioned earlier, it's a very interesting time to see how management of teams, management of these workflows sits and every time you move that slider as well, like, is it a brand new workflow or is it the same workflow with a bit of automation? What's the rules attached to it? If I start using AI more versus using it less, and it throws up many, many questions everywhere, but I think it's very much tackling some of these workflows from the ground up, but you having the foundation and the scaffolding of all the governance, permissioning and important audit as well. So then if the regulator ever does come knocking, you have got the ability to say, look, we have all of these things in place, but in the UK specifically, you have something called the senior manager regime and that basically means that whoever is the assigned SMR in an organization, their head's on the block if something goes wrong. So, and again, that's kind of a brute force way of doing it, but I think it's going to work very effectively for the short term because it means all those governance and layers and controls will get put in place because otherwise someone will go to prison if it's not, which is quite a good stick.

David Greely (24m 35s):

That's a big stick and I wanted to ask you though, because like longer term, I guess one of the questions also is if we become more and more reliant on the AI agents, but leave the responsibility ultimately with the humans, on what level that sounds very good. On another level, there is kind of the incentive issue of do the humans become somehow dulled over time? Like what I am thinking of is with autonomous driving, right? You leave more and more of the driving to the automobile, but then when it gets in a really bad problem, it throws control back to the human who probably hasn't been paying attention for the last 10, 15 miles. So the responsibility rests with the human, but then are they really in a position because they become so reliant on the machine to actually provide that oversight that they need to do? Is that a question that's out there or people thinking like wrestling with how to make sure that the humans are not only responsible, but they have the skillset and that skillset doesn't get degraded in order to oversee?

Matthew Cheung (25m 41s):

That's not a new problem. That's been a common problem in software design for a long time. When you have screens pop up and it says, I'll give flashes up a warning, don't do the whatever, click, click, click, click, click and even if you are using like Claude Code or any of those tools, every time it's doing something, it asks you, oh, is it okay if I go and do this? Okay. Is that okay to do and after a while, it's easy to slip into, oh yeah, that's fine. That's fine. That's fine and having the human in the loop there, like exactly like you say, you become dull to it and it's just becoming a pain for you. So you just, okay, okay, okay. That problem has been around for a long time. It's still not solved. That's very much. I think design is a big part of it so that people don't become dulled or bored or literally go to sleep and I don't think there's a correct approach that works for everyone for that at the moment, but that's definitely a challenge that people do and will think about but again, it depends on the, what's at stake with the particular workflow because, and again, that's what our slider piece comes in because if you are trading some huge, massive billion dollar trade, I am not going to be a sleeve of the world. I am going to be involved and the AI is going to be assisting me so I can focus on what's at present at the moment and all the other stuff around the edges, my agent can do that for me and similarly, if it's something where it's just doing post-trade booking and I just want to go, go home now because I have finished since 4.30, the bell's rung. I want to go home. I don't need to sit there filling out loads of order tickets. You might be like, I will just let the agent do it this time. So again, it depends on the scenario.

David Greely (27m 21s):

Many people think of governance as a necessary evil that you add oversight at the end of the process. You've argued that it's really something that needs to be built into the foundations of the process itself. Could you walk us through why thinking of it as a necessary evil is a problem and why you need to think about it as part of the process?

Matthew Cheung (27m 43s):

Well, since probably the financial crisis nearly 20 years ago now, there was a lot of regulation that came out after that point and that was very much that the, you had to probably work backwards and put a lot of compliance in place to satisfy the regulatory obligations that firms had. Whereas this time around with AI, how firms should be thinking about it is actually, if you have got the right foundations in place, I've got the right tooling, foundations, the right engine, right everything, I literally can go much, much faster because I know what I am allowed to do and not do and I have got all of the control mechanisms built into it. It's more, I think, think of it more like an F1 car where you are constantly fine tuning it and making it better and improving it and the better that car is, the faster and better performance you can get out of it. I think people should think about governance in that way, rather than the old mindset of looking at how regulation and compliance was kind of enforced on you and you had to comply with it. I think if you reverse the mindset and how you approach it, it means that, okay, if I have got all those tools set up, so if you think about if I am a small, scrappy startup, and you have already got people that Sam Altman said ages ago, a year ago, two years ago, about there will be a one person billion dollar company. I think there is someone who has already like half a billion, I think this year. So that's coming probably this year, I would say, because if you can use all these tools, but if I am that person who is doing a billion dollar revenue, and it's just me, obviously, I am going to be using all of these AI tools, agents, every single bit of kit that's possible out there, and I am then orchestrating all this stuff. I am then the grand overseer of it all, and then orchestrating it all. And then say the other side, if I am huge tier one bank, I can't roll out the thousand tools the one person in the garage has, I can't roll them out, because they all need to be tested. I need to check if they do comply with all these regulations, and da da da da da. So if you take those two kind of counterpoints, I think the more all of that governance layer and foundation is there, means that, okay, you're never going to have a thousand of the tools, but you might have the best 100 out of them, which is going to mean you can move quicker than the other bank down the road that's only can use Copilot, right? So it's very much that approach and mindset.

David Greely (30m 13s):

And when you look at this over time, how do you see humans and agents working together dynamically? So kind of growing together, learning from each other in a self-regulating, self-improving system?

Matthew Cheung (30m 29s):

I think what's been quite interesting, and I know it's happened to me and happened to a lot of people I talk to, where you're actually doing way more work now than you were doing before December of last year, because you are using all these tools and you have been using all these agents. And it's quite interesting where at any given time, you might have anything from five to 10 agents running around doing things for you and helping you with your work. So you are getting more stuff done, but then the way you have to manage those workflows and processes, even me myself, when I have all these things running, you have to then context switch constantly, which is not good for doing certain types of work, right? If you are doing deep work, and when I do deep work, I just switch everything off, I move to a different room where there's nothing in there and I just can sit down and do something. So you have this orchestration of all these different agents doing stuff. I think some of the stuff like OpenAI is released with Codex is starting to kind of tackle some of that. Some of the other platforms are doing similar where you will then manage, you imagine if you are using kind of JIRA with different bits of workflows you are doing, you will have like a live real-time JIRA with all these agents, and you are managing what they are doing and you are already seeing some technology come out like that. So I think that's something which is already a problem, and there's some solutions coming on to market with it and again, it depends on the types of work that you're doing. Is it deep work? Is it different types of tasks? The self-regulating, self-improving piece of it, because again, if you set this little agent off in the wild and it goes off and forks itself and starts doing other things and you come back like the next day and you are like, look, this is not what I started off with. That could be a really good thing, which sometimes you get those big aha moments quite a lot when you are using agents to do some of these tasks. And similarly, sometimes you just say to it, don't you show your screen, it's like, how stupid are you? You have got 100 billion parameters trained on every single piece of text in the world and you don't even know the answer to that. That's why you can't have some things running around kind of doing completely things autonomy and that's why it's still going to be a while yet in regulated markets until that is happening at scale. But there is other stuff where coding is a big one because it's text, it's all text, writing documents and PowerPoints and all that type of stuff. That's all changing and that's where there is a lot of big gains at the moment and you can't mess with the coding, but apart from the coding, but we are doing documents and so on. The risk of something huge is not the same as if you were putting a massive trading position on where you could be instantly owing someone hundreds of millions or billions of dollars in collateral, for example.

David Greely (33m 17s):

Along this lines, financial markets, trading desk, it's like very specific knowledge and skill sets and I think a lot of people are used to working with LLMs that have been trained, pre-trained on a broad set of data. How do you see kind of educating these agents or systems in specific jobs that would be involved around a trading desk?

Matthew Cheung (33m 43s):

It's a good question and a big problem because every single instrument asset class region has their own way of talking about the particular thing that they are trading and even in the US people on the West Coast who are trading certain commodities talk about things slightly different to the East Coast and West Coast and how you quote something in China is not the same how you quote it in London. So what's kind of happened is you have these kind of local syntaxes and nomenclature that's been formed around particular regions and markets and that's why what's very interesting is in OTC markets, which is mostly fixed income and commodities, everything is done in chat. If you think about, everyone knows the New York Stock Exchange and ringing the bell and the SPOOs and all the big indices and so on. They are highly electrified markets, which have been that way for decades. Then you filter all the way through to then fixed income on commodities where some of it is kind of digitized and electrified, but not that much of it and then so the lion's share of it, and if you look at volume, trading volume across the planet, still half of it is done when people are typing into chat, which is crazy. But that's because the nature of what they are trading can be very complex, can be nonstandard. So there is no trading venue because it just doesn't understand how to interpret this unstructured conversation that you are trying to feed into it. So I think this idea of then having the tools now to digitize that conversation, so you have got kind of the unstructured conversation, you can extract out data. So now you have turned it into data. Now once you turn it into data, ultimately, if you can start doing this kind of process, you can digitize it first, and then you can start electrifying markets as well. Because now if I have got all these different chats and I can have data that's in a GUI that I can see as a human or I can serve it up as an API, all of a sudden, you can start electrifying markets that are completely not electronic at the moment, which means that there's so many opportunities out there. And then going back to those actual conversations as well, I have been talking about data a lot, but you think about what's in a chat that's sitting on a trader's salesperson, a broker's desktop. There is really only three things in any of those chats. You have got talking about something that's quantitative or like a number, prices, quotes, RFQs, axes, runs, confirms, and so on. You have got talking about something that's more qualitative or narrative, so market color, research, commentary, news, rumors and then you have got the chit chat, the relationship, you are having a conversation with someone, what did you do last night, and so on. So those three different things require different ways to extract out that information and provide intelligence and signals on it. And that, again, varies depending on where you are.

The bit that's actually very common across doesn't matter what chat platform, what region you are in, apart from the language itself, which is easy to translate now, the narrative bit's actually pretty easy because large language models are trained on text. But when it comes to the numbers, our approach, like I mentioned earlier, is this idea of having this chat data mining tool, so that as you are starting off to use this tool in a new market, there is only a small amount of human-in-the-loop validation and verification you need to do to jump from. So when you use a foundational model out of the box, I could take like a brand new market that we haven't touched before, like Brazilian corporate bonds. We don't do anything there at the moment. You could take a whole bunch of chat history from some salespeople or traders. We can take that and just with 50 lines or so of verifying different messages in there, the accuracy before, if you are just using a foundational model, taking this example, would be anything from 50 to 65%, depending on how standard it is, because it's not been trained on that data, because that data is private to the salesperson or the trader, perhaps the chat platform itself, but it's never been used for training anything. You do that validation, you then jump from 50, 65% up to 85% accuracy. By the time you have done it a few hundred lines, you get up to 90, 95, and as you use it more, it jumps higher and higher. So someone like Coke Industries, they are a big player in US physical natural gas, they have been using tooling like this, and then their accuracy is 99% or so, and they have had 90,000 trades going through a service like this, but that's in this world of US physical natural gas and then so how we have approached the problem is we have built the tooling, so you can just use the same tooling and apply it to any market, and it will learn. You mentioned like self-learning before, this is the self-learning system where the human is helping to improve the system.

David Greely (38m 55s):

Thanks so much, Matt, for a really great conversation. There is so much to think about. Before you go, we have covered a lot of topics, and I was hoping you could kind of pull this all together, given your decades experience, to give us your vision of what the trading desk of the future might look like?

Matthew Cheung (39m 14s):

I think in the world that we operate in, which is kind of very much the OTC kind of world, means that you are going to see a lot of low-hanging fruit and easy wins, providing you have got all of that foundation and scaffolding there in the next year or so. So there is still people using spreadsheets, there is still people typing things with their fingers, so all of that will slowly evolve where you don't need to be doing these manual processes and productivity is always kind of the first place you start with a lot of these types of workflows. We are already seeing that now, as you then start to use these tools and having access to data that wasn't available before. So if you imagine, if I am a trader and I am sending out RFQs to dealers to get a price on something, when they come back with a price, I just trade with the person who gave me the best price and then all of that chatter that we had before just disappears into the ether, doesn't get used. Now, you have got the ability to use all of that data for signals and analytics. So next time I send out an RFQ, my agent is going to sit there and say, oh, last time you sent out this RFQ, this dealer gave you the best price, and this dealer was the quickest. However, sometimes this other dealer responds, and you should always go with them, because when they do respond, they give you the best price and you have the ability to start building out these data and analytics because you can tap into chat that just wasn't available before. So I think you will see that now in the short-term period. I think the medium-term period is really exciting to see how some of these markets will ultimately start to get electrified and then I think the piece that is very interesting is as agents start to get used more, and this again is some of the problems that are being solved around the identity of those agents and how the permissioning of that works, what they can connect to and ultimately, if you go forward to, I keep saying that no one is really using those for trading workflows at the moment, they will be probably sooner than we think. But by the end of the decade, no doubt there's going to be agents who have traded with agents, and how does that work and I think that's going to be exciting and probably a little bit scary as well if you haven't got all of those governance layers. What was that old 1980s film, War Games, you remember? You don't want War Games with agents trading trillions of dollars of volume of stuff because they would just bring the whole system down and again, that's where all of those governance controls are really, really important.

David Greely (41m 49s):

Thanks again to Matthew Cheung, CEO at ipushpull. We hope you enjoyed the episode. This concludes our series, How to Raise Your Agent. We will be back next week with our Summer Playlist 2026. We hope you will join us.

Announcer (42m 04s):

This episode is brought to you in part by Abaxx Technologies. AI agents are fundamentally changing how people and organizations interact with computers and the internet. A race is underway to establish the path that will govern the nature of these interactions. Abaxx Technologies introduces Abaxx Labs as its center for engaging with the developer community through the release of open source software to promote a path that empowers users to maintain control of their identity, data and agents in digital spaces. Abaxx Labs' first release is the Agents++ library, a subset of Abaxx's ID++ software development kit that has been tuned for use with AI agents. Download Agents++ on GitHub today at github.com/abaxx-labs. That's github.com/abaxx-labs.

That concludes this week's episode of SmarterMarkets by Abaxx. For episode transcripts and additional episode information, including research, editorial and video content, please visit smartermarkets.media. Please help more people discover the podcast by leaving a review on Apple Podcast, Spotify, YouTube, or your favorite podcast platform. SmarterMarkets is presented for informational and entertainment purposes only. The information presented on SmarterMarkets should not be construed as investment advice. Always consult a licensed investment professional before making investment decisions. The views and opinions expressed on SmarterMarkets are those of the participants and do not necessarily reflect those of the show's hosts or producer. SmarterMarkets, its hosts, guests, employees, and producer, Abaxx Technologies, shall not be held liable for losses resulting from investment decisions based on informational viewpoints presented on SmarterMarkets. Thank you for listening and please join us again next week.