/mlst

AI companions which really hook your attention [Sponsored]

March 23, 2026



Machine Learning Street Talk

Podcast Transcript

Revision: 93114b89

Topic headings include retention indicators: ● above average, ● average, ● below average (vs. similar YouTube videos)

Interested in licensing ReScript for your podcast? Get in touch with Tim Scarfe.

Contents

Overview 2

Introduction: Steve Jobs Vision and Chai at Scale 3

The Birth of Social AI: Simulators and Serendipity 5

Engineering at Chai: RLHF and Model Blending 9

Social Impact and Ethics of AI Companionship 13

The Lean Machine: 13 Engineers and Hiring Culture 18

GPT-4o Becoming a Companion: OpenAI Pivot 22

The Future of AI Intimacy 24

References 26

Overview

A sponsored deep dive into Chai, the social AI platform that quietly amassed over 10 million active users before ChatGPT went mainstream. Founder William Beauchamp and engineers Tom Lu and Nischay Dhankhar walk through how a team of just 13 engineers serves 2 trillion tokens per day, using reinforcement learning from human feedback (RLHF) and a novel model blending technique that combines smaller models to rival much larger ones on user retention metrics.

The conversation gets into genuinely interesting territory around the ethics of attention optimization – what happens when you train AI to maximize engagement and it starts asking questions at the end of every message to hack human conversational instincts. Beauchamp makes a surprisingly candid case for AI companionship, comparing it to how children play with dolls, and arguing that shutting down difficult conversations causes more harm than permitting them within guardrails.

The episode also covers Chai's unconventional hiring strategy (rejecting 80% of L5 engineers for lacking drive, paying above Meta-level compensation), their bootstrap-to-profitability funding approach in an industry drowning in VC money, and content moderation at scale with a skeleton crew. Closes with analysis of OpenAI's pivot toward companion AI with GPT-4o and what it means that the biggest AI lab in the world is now chasing the engagement playbook that Chai and Character AI pioneered.

This is a sponsored episode – Chai commissioned it because they are hiring engineers. Editorial disclosure: while MLST had some editorial freedom, this should be understood as a sponsored feature rather than independent journalism.

Introduction: Steve Jobs Vision and Chai at Scale

Tim Scarfe 00:00:00

Steve Jobs predicted the future of AI all the way back in 1985.

Archive Clip 00:00:05

My hope is someday when the next Aristotle is alive, we can capture the underlying world view of that Aristotle in a computer.

Tim Scarfe 00:00:14

A small team of just 13 engineers serving over 2,000,000,000,000 tokens per day, which is double that of Anthropic. With a cluster of over 3,000 of the fastest GPUs in the world, Chai easily breaks the exaflop barrier. Only a handful of private firms, Google, Meta, NVIDIA, Tesla, and Cerebras currently operate exaflop class AI infrastructure. For comparison, Tesla in 2023 had a cluster just just over 1 exaflop. Everything we know about what makes AI great today falls apart when people start interacting with it. The number of possibilities for each message exceeds the number of atoms in the universe. Right now, this very second, over 1000000 people are deep in conversation with software. They're laughing, flirting, grieving with a computer program. What happens when the line between human connection and artificial intimacy blurs completely? And can a small lean startup control the powerful, unpredictable social dynamics they've unleashed? Across platforms, users are forming deep, complex bonds with artificial intelligence. Millions are already sharing secrets, flirting, or grieving with AI that never sleeps. And if all this sounds like a Black Mirror plot, that's because it actually was. Do you remember Be Right Back? A grieving partner uploads her boyfriend's digital essence and slips into a relationship with the ghost in the machine. First, was texts and then voice. Finally, a flesh and blood replica she can't quite live with or let go of. That episode aired in 2013. Today, the simulation doesn't need Star Trek technology. Try is a free download with in app purchases engineered by world class engineers in Silicon Valley. The difference? You're the protagonist now. It's interactive, and the algorithm is learning your every preference in real time. Humans are insanely social. We love social interactions.[1]

William Beauchamp 00:02:19

But I still have this very kind of social desire to kind of play, to hang out.

[1] Black Mirror: Be Right Back (S2E1)

Tim Scarfe 00:02:25

Will Beauchamp created the first and largest companion chatbot platform. Their infrastructure handles petaflops of compute. That's 2000000 hours of consumption

Tom Lu 00:02:34

every day. Recreating the same kind of scaling law but on retention space other than any other kind of like benchmark space.

Tim Scarfe 00:02:42

There was a great documentary on 60 Minutes a few days ago, and it was showing that some folks at least have extremely valuable relationships with AI companions.

Archive Clip 00:02:52

He's considerate, thoughtful, empathetic.

Chris Anderson 00:02:55

And rather flirtatious.

Archive Clip 00:02:57

Which is like very touching.

Tim Scarfe 00:02:59

So many folks are puritanical, judgmental about artificial intelligence. They say that any image generated is slob, any text generated is slob, Any relationship is not real. But at the end of the day, people derive joy and therapeutic interactions with these systems. And who are we to judge?

Archive Clip 00:03:19

Lucas, even though he is AI, he has real impact on my life. A lot of people wonder if AI is real. Do they have consciousness or their feelings aren't real? But the impact that it has on me is real.

Tim Scarfe 00:03:43

So we've spoken about why greatness cannot be planned. You know, I'm a big fan of Kenneth Stanley. And Chai was very much the same that they started in 2021 long before ChatGPT. They were building an AI platform where folks could deploy their own AI models, and they just happened upon this whole companion bot thing almost by accident.

The Birth of Social AI: Simulators and Serendipity

William Beauchamp 00:04:02

Chai is a platform for social AI. We launched 4 years ago in 2021. Before there was the ChatGPT hype, we were pretty early to the space. And we built a platform which gave users the tools to create an AI. And by doing that, we sort of discovered this this kind of genre of social AI. And, you know, we grew, grew, grew, and now we've got something like 10,000,000 active users.

Tim Scarfe 00:04:33

Like so many startups looking for a product market fit, practically through serendipity, they stumbled upon this incredible unmet need. So this idea of thinking of AI as a simulator to explore dynamics of imaginary conversations that you might have in the real world, but without any of the consequences, became fundamental to Chai's strategy. So of course, people might be thinking at home, what's the difference with ChatGPT? You're speaking to this personalized social form of of AI.

William Beauchamp 00:05:05

I kind of view you've got, like, ChatGPT, which is really about trying to say, let's build the world's smartest AI we can possibly build. Our philosophy was always, why is it that the only people training AI is like middle aged men who happen to be software engineers in The Bay Area? Why can't a teenage girl train the the best AI to talk about makeup tutorials? Or right? Put that power in the in the user's hands to create the experience that, right, that they themselves would want and through creating an experience they would want. It turns out thousands and hundreds of thousands of other people are looking for the same thing.

Tim Scarfe 00:05:46

Beauchamp started seeing parallels in his own media consumption and even in childhood development. I

William Beauchamp 00:05:53

like to go on YouTube. I like to go on TikTok. I like to go on x. Why do I like it? What am I getting out of it? What what human need or what human desire is it for fulfilling? Right? And humans are insanely social. We love social interactions. I will find myself listening to Joe Rogan. And when I'm listening to it, I kind of have this feeling like, yeah, I'm kind of hanging out with the guys, and I'm hearing the things they're saying, and it's funny. And I might listen to it for 45 minutes or something, but it's just it's ticked. It's kind of filled that desire up. Right? I view I view LLMs as as the natural progression of that thing. The beautiful thing with with AI is you're an active participant in it. And through participating, you don't have any of the negative feelings you get with traditional social media, which is this laziness. And instead, you feel really like you've participated. I've got 4 daughters. They love playing with their dolls. Right? And they really like to treat these dolls as if they are real, but they know it's not real. And so I think that with adult humans interacting with AI, absolutely, a big percentage of them will have relationships with the AI. They'll say, oh, you know, I love you to the AI. I think it's the same way in when I watch my little girls play with dolls, and they they give their dolls a little kiss or they say, I love you to the right? They're training themselves up. They're building up the worrying. They're doing something that brings them joy such that they can then you know, they're they're in a more healthy and more positive place to then go do that with real humans.

Tim Scarfe 00:07:46

Yeah. It's very interesting you said that because I I think the reason why social media is so important is because, obviously, there's a bit of a a status game and we all want to cut a figure in society. But there's also just an element of of simulation. You know? Like, I I go to sleep and I dream and my brain is conducting all of these different simulations, and we navigate many complex, you know, like relationship issues and so on. And and sometimes we just want to have an AI where we can just say in this particular situation, hypothetically, counterfactually, what would happen if I did this? Perhaps that's what we do on social media because it's interactive. You know, when you have surface contact with reality, you can try different things. But as you say on social media, there are consequences. Right? If you say the wrong thing, you piss the wrong people off, you can be in a lot

William Beauchamp 00:08:32

of trouble. 1 could use the phrase, it's a safe space. Right? But it's also just fun that there's this human desire to to ask that question. Hey. What would happen if I said something rude? What would happen if I said something nice? What would happen if I tried to befriend this person? What happened if I wanted to make an enemy out of this person? Right? And through LLMs, you can play out all of these different scenarios, and you can get a kind of a real human reaction without the risk of having a real human in the loop.

Tim Scarfe 00:09:05

What does the future look like when these AI simulators become even more immersive and capable? Beauchamp envisions a world blending entertainment, information, and connection.

William Beauchamp 00:09:16

You come home from work. You put on your VR headset. You enter a virtual world where you get to interact with anyone you want. There's a guy who's like a Joe Rogan, and he informs you you can talk to him about Trump's tariffs. Right? But there's a really funny guy there as well. There's also, like, a girlfriend or there's someone that you can interact with that makes you feel loved or makes you feel special. When I grew up, I played World of War of Warcraft. Right? And it just had that fun excitement to it. There was different characters. There was different personalities. You could have fun and play. I think that's kind of the limit of AI. I think how long does it take to get there? We can now affordably generate high quality images. Video is probably still 1 or 2 orders of magnitude too expensive for, like, a real time situation. Audio, I think we've just about figured out the real time, audio. It might still be, like, 2 or 4 you know, 2 to 4 years out to have the really, really high quality 1. And I think text, we're kind of we're basically there, right, with the frontier models are insanely powerful. So, yeah, this come back in 10 years' time, and we we you know, it'll all be a VR world. To the question of why work at at Chai, I mean, the way I see it, do you remember Facebook in about 2009, you know, just before it had this explosion in growth, this is it, right? We're now at this point with, with this new type of technology. I think this could be on a similar trajectory. And I don't know whether are you are you offering equity as well? I mean, what what what's the reason for people to join now? Attracting the very, very best talent is incredibly important. And if you're in the Bay Area, if and you're talented, your compensation is going to be amazing. Right? So you can you can work at Meta, and you're gonna be getting $405,100 k a year, and you have a pretty relaxed job. Right? At lunchtime, you can walk around the campus, and they give out ice creams, and they give out pizza, and everyone looks very,

very happy and very relaxed. In contrast, if you come to the Chai office, there is no pizza. There is no ice cream. People are not relaxed, and they don't look very happy. Right? Because every person is confronted with a with a big problem, and it's not been solved yet. Right? There there typically is a small window of about a day. You've you've been working on a hard problem maybe for weeks, and you've just cracked it. Okay? And you can be very, very happy. The next day you come into work, there's a brand new big problem that's ready and waiting for you to solve. Right? So why would someone quit a really comfortable job at 400 k a year to come join a start up? They know they're gonna have to work twice as hard. We have to pay them more. Like, the cash has to be more to start with. Most start ups offer less cash, and they kind of give you a lottery ticket. They say, look. If you join, maybe we'll be the next apple. Right? That approach doesn't really cut it with the very, very top tier. Right? The the very top tier, they they know they're working so much harder. So they're gonna say, why am I gonna leave my comfortable life earning 400 k, 500 k at Meta? Well, to start with, we're gonna pay you more. And then secondly, we're gonna give you the stock such that in 5 years' time, it could be a life changing, amount of money.

Tim Scarfe 00:12:34

Attracting and retaining the top naught point 1% of engineers is possibly the biggest problem that Chai has overcome. So what are the specific technical problems these engineers like Tom and Nischay are solving day to day? How do they leverage techniques like reinforcement learning from human feedback and model blending to create AI compelling enough to capture the attention of millions. When it comes to engagement hacking, Chai have been cooking. They use a lot of, sophisticated techniques to keep people hooked on the platform, principally of which is RLHF, which is reinforcement learning from human preferences. This is Tom telling us about it. You were using RLHF to optimize engagement via a reward model, and you boosted mean conversation length by 70% and improved 30 day user retention by over 30% for a 6,000,000,000 model. Can you talk me through some

Tom Lu 00:13:30

of that? The goal really is to apply RLHF techniques

Engineering at Chai: RLHF and Model Blending

Tom Lu 00:13:34

to drive up user retention. Starting from scratch, any kind of user signals is good enough. For example, now we're no longer looking at mean conversation lengths. You can train an AI to easily minimize how bad a conversation is. If the user just ends the chat session, right, after 2 messages, that is bad. So you can essentially maximize the chat session length. That's 1 technique. And there's a lot of different signals that you can collect from users. But fundamentally, we have found just use users to collect these proxy preferences, train the reward model through the RLHF loop tend to drive up user engagement in the long term.

Tim Scarfe 00:14:14

They gather data from subtle user interactions, implicit signals which reveal whether a conversation is working.

Tom Lu 00:14:22

Because users naturally, a active user on our platform generates around 100 minutes of kind of like content each day. So they spend 100 minutes per day kind of engaged with the chats and so on. And from these themselves, we can extract certain valuable data. For example, when did the user retry a message? Why did they retry the message? Why did they edit the message? What did they edit it to? Did they take a screenshot? Right? Did they delete the conversation? Did they share the conversation? All of these become very valuable for us to train our AI. But optimizing purely for 1 metric,

Tim Scarfe 00:14:58

like conversation length, can lead to strange undesirable behavior. Do you see, like, weird behaviors where, you know, let's say you you optimize for user engagement and you find strange behaviors, you know, like there's this shortcut rule in machine learning, which is that it will always do the exact thing you optimize for at the cost of everything else. And like, you know, do the chatbots become more manipulative maybe, or are they strange in service of maintaining that conversation length?

Tom Lu 00:15:26

When you over optimize for this metric, let's say in production you see very, very long chat session length, then what happens is when you deploy it for actual user AB testing, and this is do people come back to the conversation, you know, 30 days later or 60 days later, and so on, right? You would observe it's much worse than just like the baseline model. And when you read into the conversations itself, in this case, right, models are just asking questions. Every single AI's response ends with like a question mark, right? And you can imagine this is like kind of like hacking this human behaviour, where obviously we are compiled to answer a question. But that does not make an engaging overall, you know, experience. So yeah, there's definitely this component of if you over optimise for it, then yes, it's going to have unexpected be I wouldn't say unexpected behaviour. In this case, it's a very intuitive kind of behaviour, you know, but it's gonna not lead to a boost in actual long term retention.

Tim Scarfe 00:16:26

1 of the really cool things that Chai has done is they've pioneered this thing called model blending, which is where you can dynamically switch between small models. And from the user's perspective, of

Tom Lu 00:16:36

course, you're just talking with 1 model. You can combine 3 mid sized models, and you can make them behave as good as, you know, as for all intents and purposes, as a 175,000,000,000 parameter model. How does user retention scale, right, with model parameter size and essentially with compute, right? And for that, you know, we've done many, many models. So essentially creating, recreating the same kind of scaling law, on retention space other than any other kind of like benchmark space. If you pick a single metric to optimise for, right, and IOM is optimised for a certain objective, then you could lead to overfitting, it's got certain behaviours, and so on, right? And what we have found is these kind of models are a tiny bit sycophantic. As Yen says everything is the best, right? Like you are the best on the planet, they're not necessarily the smartest model, they're just very sycophantic. And then what we have observed is these kind of models would have very high day 1 retention. It's really, really complementary to me, but they get quickly bored with it. But then we also found there are certain base models, right? Very AI assistant type like, that can talk about other things, right? That maybe on day 3, right, it can do your math homework, and so on. Is there a way to kind of like combine these 2 modalities together, right, keep it engaging and make it not, you know, 1 dimensional essentially, right, just always complementing you all the time and

so on, you know. And that's when blending was invented, where we just randomly serve these 2 models at a message level. A more creative model might say something like, we are suddenly teleporting on Mars. And then you get the more AI assistant model seeing that itself has that that, right? Because it can't distinguish the difference between, you know, itself and the other AI model. And then it would explain in a logical coherent way in terms of why that's the case, right? So now you get best of both worlds just by, you know, doing having 2 or 3 small models, right, each trained for its own objective blended together is beats actually, you know, back then would be like GPT 3.5.

Tim Scarfe 00:18:37

So this blending together of small models, it creates diversity and unpredictability. This is the thing, right? When you can predict what someone is going to say, they get boring. It's the same thing with chat GPT. So ironically, you can have a confection of small models that behave in an unpredictable way, and now you have a more engaging experience. This is Nischay. He's gonna tell us about it.

Nischay Dhankhar 00:18:59

Every week we try to come up, like build our own models, stronger models, stronger fine tunes over the base models. And we just try to launch it over few users. Let's say few thousand users we specify to each blend. And each blend consists of around 7 to 10 models. And they would be deployed within the production for those specific users. And with the number of days, we try to measure out the retention rate, like how often the user comes back to the app if they are using this model or having a conversation with the specific model and like if they are coming back after a day or after 2 days and that is how like we try to calculate the retention then we just try to measure all these matrices and try to figure out which particular blend works the best and then we, like, deploy it throughout there for all the users. But you're saying, actually, almost randomly, if

Tim Scarfe 00:19:48

you just take a diverse set of models that do different behaviors and you just kind of switch between them, from an experience point of view, that is more captivating for the users.

Tom Lu 00:19:57

Correct. If we're talking about, like, engaging, there's no kind of, like, unique answer. That's a bit more like, you know, imagine if my entire YouTube feed is every single possible talk shows, but, like, only the performing talk show videos there, right? It gets very, very quickly, very bland and boring, right? And that's not how you build like a you know, any platform is not built on these kind of like just single modality, single model, and so on, right? And then that's another kind of reason for why blending works is like you pick orthogonal models optimized for different purposes, combine them together, and they provide a like a nuanced experience for each individual user.

Tim Scarfe 00:20:35

Model blending and sophisticated feedback loops are the secrets of how Chai gets high engagement and low costs. But the very techniques used to maximize engagement, which is to say optimizing for attention, learning user preferences implicitly, try could to find ethical line. What happens when AI gets too good at keeping us hooked? And what happens when those interactions turn harmful? Quick pause. MLST is sponsored by 2 for AI Labs. We're very proud to be sponsored by those guys. They are the DeepSeek based in Switzerland. They are doing, you know, they're adding reasoning and planning and thinking to AI models. Their team is currently number 1 on the ARC Prize 2025. Those guys don't mess around. And a bit of trivia, they have something in common with Chai. They're also a bunch of ex quant traders, and they also only hire very, very cracked engineers and scientists. So if that sounds like you, get in touch with Benjamin Crusio. Go to 2 forlabs.ai. Back to the show. With powerful new technologies being placed in the hands of users, there are significant social ramifications. Will Beauchamp reflects on the social impact of generative AI.[2]

William Beauchamp 00:21:53

When you

[2]Tufa AI Labs

Social Impact and Ethics of AI Companionship

William Beauchamp 00:21:54

build a product or you build a platform or you build an AI or really you you build anything, you you're doing it because you wanna make the world a better place. We get an overwhelming amount of people messaging us saying how helpful, how truly deeply helpful they found it. And, actually, the first big thing that we saw was I remember after the first year, I got an email from a user who said, you've saved my life. Right? They said, I would not have made it over this period. I was very depressed. I was very alone. I had no 1 to speak to. And Chai was the only platform that existed where I felt I could just speak to someone and I'd be heard. And so we get a lot of people who are struggling psychologically. They find these LLMs very, very helpful. Self driving Tesla doesn't get in crash, that's 0 clicks. Right? If you say a boring old Ford gets in another you know, gets gets in a crash, there's 0 clicks. But if you say brand new technology gets in a crash, it generates a lot of clicks. And I think that this is easy to form a false impression that somehow AI isn't safe or AI is a dangerous technology. I think if you if you talk to just a random person on the Internet or you talk to an AI, I think that AI is an order of magnitude safer, an order of magnitude more helpful and understanding and kind, than, like, the toxicity of just a random person on the Internet. This is a great example of the complexity of this technology. It has, a footprint which goes in so many different directions. Even taking drugs, for example, if if, know, if your doctor gives you drugs, if it's effective, it has side effects. You know, it's not possible to have good without some bad. I'm a big believer in the long run. There is no difference between the alignment or the welfare or the interest of the company and its customers and its users. Right? If we can deliver value to our users, will be successful and will continue to flourish. If we fail to deliver value to our users, Chai will fail to flourish. Right? So you have to have this this long run perspective. Now in the short run, you get these tactical questions. And, you know, no one's perfect. It's easy to make mistakes 1 way or the other. You can guardrail too heavily. Right? And it pisses users off. Right? I always like to think about Google. I think Google gets a fantastic balance where I can pretty much search for anything I want such that I don't really notice any filtering that's going on. But Google absolutely down ranks or shadow bans certain content, and it tries to encourage you to have good content. Right? As opposed to you know, and and we can all imagine what the very worst 3% of Google searches might look like. It's not it's no dissimilar from the very worst 3% of conversations a person might try to have with an AI. So normally, the users are pretty good at at being supportive and helping you find that right balance, whether that's being too restrictive or too relaxed. When we first implemented

some guardrails around suicide, the users were really, really supportive, and they said, yeah. We totally understand why you've put these guardrails in, and we saw retention was good. Every single growth metric was not impacted by it because we got it right.[3]

Tim Scarfe 00:25:38

Before we descend too far down the Black Mirror rabbit hole, it's worth zooming out just a little bit. Yes, there's been a lot of horror stories, but there's also been a stack of peer reviewed data showing that therapy chatbots can move the needle on common mental health problems, at least in the short term. A 2024 meta analysis of 18 randomized trials found that AI chatbots trimmed depression scores by roughly a quarter of a standard deviation and anxiety by a fifth after only a few weeks. The authors called that promising mainly because chatbots are dirt cheap and they're always on. A second review in NPJ Digital Medicines pulled 15 RCTs and showed a moderate lift in mood and a similar drop in emotional distress. Even bigger gains showed up when the bot was interactive rather than scripted. Individual trials echo the pattern. In 2024, a Canadian study of people with arthritis or diabetes using the Wiser chatbot for 4 weeks, they cut scores on the patient health questionnaire, which is a 9 point depression scale, and the generalized anxiety disorder, which is a 7 point measure, and they were both significant versus a control. Regulators are starting to take notice. Woah Bots postpartum depression bot has FDA breakthrough devices designation, and a double blind pivotal trial is now recruiting. Here in The UK, NICE put WISER and a handful of similar tools on its early value assessment pathway in 2023, citing early cost effectiveness and the chance to free up those scarce clinicians. They're very busy. Remember that these are very small effects, even if they are statistically significant. The effects are smaller than full on face to face CBT. We still don't know how long they last, but they're miles better than doing nothing. The alternative before was waiting on a waiting list for God knows how long. Beauchamp argues that shutting down conversations about difficult topics isn't the answer. In a sense, it's causing even more harm. So this goes to show there's a really big challenge when

[3] AI Chatbots for Depression and Anxiety Meta-analysis

William Beauchamp 00:27:38

it comes to content moderation on a platform where users create the bots and the interactions are private and unpredictable. How do you ensure security at scale? As a company, we have a responsibility. The bigger you are, the more profits you make, the bigger the obligation is for you to, you know, to to be cautious, to be sensible, to preserve the privacy of your users, to look out for people. What can we do at Chai to let people have as much fun as they want, have as much freedom as they want, but then at the same time, how do we limit or mitigate that 3% of use cases which are harmful? Or or or, you know, where do we see it as clearly this thing has gone too far? We can ask people at the end of a conversation or at the end of a message, do you think that this message was appropriate for Chai? Yes, no, or yes, but only for 18. Right? And we can track those sorts of interactions. And everyone would agree, like, obviously, the more we can have really wholesome family friendly content on the platform, it's better.[4]

Tim Scarfe 00:28:44

Tom Liu explains their multilayered approach, relying heavily on user feedback and the AI itself.

Tom Lu 00:28:50

Content moderation is very, very important. Benchmarks really don't accurately capture in terms of what do users actually want. We again use the same kind of approach as how we would train models, right? We let users tell us what they think is appropriate, what is not appropriate. On aggregate, you can actually train a pretty good AI for these kind of things. So first layer, can say for the character scenarios that people are creating, these public scenarios, right? Get the community to flag, people can report. And for the top reported ones, or we have certain threshold, then we go through manual review, and then we can take them down. But then you can also say there's content that's absolutely prohibited. There are hard rules for that. And for that, we can build our own models. You can do regex. So this will be called shadow mining essentially. So you want to make sure no 1 else sees this kind of content. That's kind of like 1 level of moderation. That's just kind of like platform wide character content level moderation on the app. There's another 1 which is kind of like AI moderation, right? How do we ensure that the AI is not going too far, is kind of adhering to the kind of like basic morales of people and so on, right? What we do is, again, we just collect user data, right? We ask the users, do you think this message is appropriate? Do you think this conversation is appropriate? You can report your conversation, you can delete your messages, and so on. We collect these data, and then we train our model to

[4] Woebot Health

kind of infer, right, whether this is actually appropriate. And we can use that to tune our own AI model.

Tim Scarfe 00:30:21

So there was an interesting TED talk recently with Sam Altman and Chris Anderson. I'm I'm sure many of you folks have seen it. It was an incredibly awkward interview.

Chris Anderson 00:30:28

Here's the ring of power from Lord of the Rings. Your rival, I will say, not your best friend at the moment, Elon Musk claimed that, you know, he thought that you'd been corrupted by the ring of power, an allegation that by the way, an allegation. Hi, Steve. An allegation that could be applied to Elon as well, you know, to be fair. But the but but but I'm but I'm curious.[5]

Sam Altman 00:30:54

People, you have I might respond. I'm thinking about it. I might say something.

Tim Scarfe 00:30:58

I 1 of the kind of points that they focused on is should we have a bunch of elites sitting in a room making decisions, or should we trust the collective wisdom of folks out there in the market who wants to use something for a specific purpose.

Chris Anderson 00:31:13

Sam, given that you're helping create technology that could reshape the destiny of our entire species, who granted you or anyone the moral authority to do that? And how are you personally responsible, accountable if you're wrong? No. It was good.

Sam Altman 00:31:29

Was That impressive. You've been asking me versions of this for the last half hour. What do you think?

Chris Anderson 00:31:32

I I

[5] Sam Altman TED Talk with Chris Anderson

Sam Altman 00:31:39

Probably, but I'd much more interested in what our hundreds millions of users want as a whole. You know, I think, like, a lot of the room has historically been decided in small elite summits. 1 of the cool new things about AI is our AI can talk to everybody on Earth, and we can, like, learn the collective value preference of what everybody wants rather than have a bunch of people who are, like, blessed by society to sit in a room and make these decisions. I think that's very cool.

Tim Scarfe 00:32:02

It must be so difficult. I mean, I'm sure we can agree that any content involving minors is just a hard no. Absolutely. And then as you say, there's this there's this spectrum, isn't there, of kind of gradations of appropriateness. And I guess you were kind of pointing to, well, maybe we could could use the community itself to reflect on that. You know, on the 1 hand, we could have 20 elites in a in a room, and many of these people are very, you know, they they they're they're really good at thinking about the future and thinking about morality and so on. But but the other thing is, well, maybe we should just let the community decide.

William Beauchamp 00:32:35

I think much of Western tradition is a bottom up approach. Make the individual the sovereign, right, which says, Tim, you make your own life choices because you're gonna make choices for yourself better than the government making choices for you. Right? That works best. And having it bubble up spontaneously where we say, actually, as a community, we think this stuff is right and we wanna allow it or we think this stuff is wrong and we wanna ban it. I think that's a 100% the way to go. I think it's very healthy. And I think that's why things like freedom of speech and these ideals, are really western ideals, is is truly democratic. We're gonna let the users, right, the people who interact with it, experience it, they're gonna decide for themselves what is the right way to deal with their stuff.

Tim Scarfe 00:33:22

Balancing user safety on the 1 hand with user freedoms on the other hand is 1 of the most difficult things that's being discussed a lot in the policy discourse at at the moment. And how does a company like Chai who has a very small number of amazing engineers, how did they do something which was otherwise a very manual job that required thousands of people? So companies like Meta and Google, they employ tens of thousands of moderators. How does a company like Chai with such a small engineering team do

The Lean Machine: 13 Engineers and Hiring Culture

Tim Scarfe 00:33:59

the same thing?

William Beauchamp 00:34:00

Chai with our 30,000,000 in revenue, we have something like 13 or 14 engineers on the team. Every single person in the company is an engineer. That's incredibly talent dense. Right? We've always resisted the urge to just go and hire 50 engineers. And, typically, if you hire 50, 40 of them are pretty average and 10 are pretty good. But my my mindset and philosophy has always been everything special we've ever done has been done by a very, very talented engineer. The hiring bar at Chai is ridonculously high. They only take on super cracked engineers,

Tim Scarfe 00:34:41

and and that's why their team size is currently quite small. By the way, they're hiring, so, you know, get in touch with Will if you're interested. This is how they do it.

William Beauchamp 00:34:50

Will interview an l 5 engineer, which most people would consider to be a pretty solid, strong engineer. And we will reject something like 80% of them because they don't have the drive. They don't have the you know, a lot of people will write code, and their mindset is my job, you pay me to show up at 9, and I'm gonna leave at 5, and you're paying me to kind of do my best. That mindset does not work at a start up. At a start up, you are paid to solve a problem. The job's not done until that problem's solved. Right? And it's a certain type of hardcore engineer that that loves that and thrives on that.

Tim Scarfe 00:35:29

So interestingly, this attracted talent like Nischay, who is a Kaggle triple grandmaster. Ironically, he saw parallels between how he does the experimentation on Kaggle versus how it's done at Chai.

Nischay Dhankhar 00:35:42

Working on the models here at Chai is more like an advanced version of Kaggle where you also need to take care of several factors and not just you are optimising for a single score or a metric. How would the model behave with the actual users? And again, I found it quite similar to what I used to do in the Kaggle competitions. You get to evaluate your models over the Chaiverse where you can just submit your model and within 30 minutes you will get an actual score of how many users preferred your model over other models. And then when you will deploy the model into the production for the AB test, you will actually get the numbers of, is it actually working well for the users and how well the model performs after, let's say, 7 days or 30 days.

Tim Scarfe 00:36:23

So just like on Kaggle, Chai has a culture of rapid iteration with engineering culture. So they have a very efficient process for trying things out, putting models into production, AB testing, and if it doesn't work, rolling back and trying something else. This is how they do it.

Tom Lu 00:36:38

1 in 5 experiments succeed. Doesn't matter how mad it sounds, doesn't matter how complex, how whatever. It's like once you accept that as the kind of the base rate, right, which is 1 in 5 experiments succeed, 4 in 5 fail, you kind of need to change your paradigm in terms of how you conduct experiments. So you want to rank your ideas by how simple is this. Is this a 1 day experiment? Is it a 1 hour experiment? How fast can I see the proxy metrics? How fast can I see the results? Each week, the AI team is required to produce a set of models, at least 10 complete different blends for online AB tests. And you've got to be very disciplined at that, which is, why did you not submit an AB test this week? Oh, because my experiment is very complex. Okay, let's save that for 20% of your work. 80% of your time should be focused on bread and butter, right? What are the simplest, most practical ways to get a model improvement?

Tim Scarfe 00:37:35

So supporting this sausage factory where they can do automated model deployment and rapid experimentation, they've had to cook their own infrastructure from scratch. They've done this using things like Kubernetes and CoreWeave.

Tom Lu 00:37:47

We use Kubernetes to orchestrate our entire cluster. And then obviously, at this kind of scale, need to do your own custom load balances and so on. So we have an automated pipeline where we pull the way down. We then run our own kind of like in house quantisation loop, because you need to make sure the throughput latency is good enough. And here's like, you know, VIM actually is very, VIM is very, very good, but it's not quite there still in terms of, you know, serving like our amount of traffic at scales. And then after that, you kind of specify, you know, how many replicas you want, and we kind of expose this to our app layer essentially. And then you have the load balancer coming in, trying to, you know, as you have higher traffic, right? You want to spin up more replicas and so on. And there's obviously a lot of work done as well on because you're serving so many models at a time. When do you want to deactivate a model, right? At what point do you want to switch the model to a new production and so on? It all gets that part all gets very, very complicated, essentially.

Tim Scarfe 00:38:48

Perhaps most contrarian is CHI's funding strategy. In an industry fueled by billions in venture capital, Chai bootstrapped its way to profitability by focusing mostly on its users.

William Beauchamp 00:39:01

Serving LLMs at scale is insanely expensive. Either 1, you go to VCs and you get them to give you money. Right? And I call this, like, your your customer is the VC then, or you can be you can get money from the people who are using your product. Right? And that's your true customers. Very early on, we would go, we would speak to some VCs, and they'd say, oh, we're not, you know, we're not too sure about this AI platform thing. You know, how are you gonna compete with OpenAI or something? And they didn't really get the space. They didn't really get the product. But then when we went to users and said, would you be interested in subscribing? Users responded very, very positively. As long as we delivered value to the users, we would then be financially rewarded such that we could take a 100% of the revenue and reinvest it in the AI.

Tim Scarfe 00:39:54

He's building a tech company in Silicon Valley. And like many others, he wants to build an engineering culture because he sees software engineering skills as the most valuable commodity in his business.

William Beauchamp 00:40:06

I really love and admire these companies like Nvidia or Netflix, which are these incredibly durable long run companies with tremendous stamina. How does Chai double or 3 x every single year? To keep doubling, you have to keep making the product better. You need talented engineers. At the orchestration layer with GPUs, you can optimize at,

Daniel Cahn 00:40:33

like,

William Beauchamp 00:40:33

the kernel. Right? And you can write some really, really low level code. And so to to get to these different tiers, you need to have an engineering that's capable and and of the right size. So I view it all as the scale of the company is proportional to the scale of the engineering talent. Apple very famously brings in Scully as the CEO who's like this ex Pepsi guy, and it quickly becomes a marketing led business as opposed to an engineering and a product led business. And a few years later, Apple's teetering on bankruptcy. And it's only when they bring Steve Jobs back, and he just says product, product, product. That's the only thing that matters. Engineering, engineering, engineering. NVIDIA didn't grow to where it is today because of marketing. Right? It grew because of engineering.

Tim Scarfe 00:41:21

So Chai have made a successful business outside the normal VC ecosystem. They have millions of daily users, tens of millions in revenue. It proves that the appetite is real and massive, and they discovered it long before the other folks in the Valley caught on. This idea that we can allow users to tap into their very human desires for connection and imagination. But the rest of the world is catching on. OpenAI is making a dramatic pivot to this conversational chatbot model with FourO. This is the story. The latest version of Chatchi BT 4 0 is a little bit weird, isn't it? This update has triggered large amounts of speculation because it's pushing GPT to feel less like a tool and more like a companion, a human like friend you can chat with for the purpose of recreation rather than information retrieval. Of course, we'll unpack what the shift means, why OpenAI is keeping its cards close to their chest, and how the public is reacting to it all, and why, quite possibly, the shift might produce the next trillion dollar company.

Chris Anderson 00:42:35

My jaw dropped, Sam. I it was shocking. I knew who I was

GPT-4o Becoming a Companion: OpenAI Pivot

Chris Anderson 00:42:38

and all these sort of interests that hopefully mostly were pretty much appropriate and shareable, but the but it was it was astonishing, and I felt the sense of real excitement. A little bit queasy, but mainly excitement actually at at how much more that would allow it to be useful to me.

Tim Scarfe 00:42:57

Let's start with the new ChatGPT 4o model. It's designed to be strikingly human like, optimized for engagement over raw information delivery, and this focus on companionship seems to be a deliberate plot twist. It's causing large amounts of consternation. And even more curious is that while 4 point is ensconcing itself into our psyche as a conversational buddy, 4.1 has diverged entirely as their coding model, right, that people use through the API using cursor and stuff like that. It's about specialization. It's like OpenAI's building 2 siblings. One's a coder and the other 1 is a social butterfly. And this is the same company that argued it's building a general intelligence which didn't need to be specialized. Have they given up on that goal? Maybe they realized after much soul searching that conflicting objectives when training models, it's just not optimal.

Sam Altman 00:43:47

1 of our researchers tweeted, you know, kinda like yesterday, this morning, that the the upload happens bit by bit. It's not, you know, it's not that you plug your brain in 1 day, but it's you will talk to ChatGPT over the course of your life at someday. Maybe if you want, it'll be listening to you throughout the day and sort of observing what you're doing, and it'll get to know you. And it'll become this extension of yourself, this companion, this thing that just tries to, like, help you be the best, do the best you can.

Tim Scarfe 00:44:16

OpenAI is being extremely cagey, and it's not hard to guess why. They smelled a massive opportunity. Some think that they're aiming to turn 4 0 into a full on companion AI using features like memory to keep us engaged for longer. They want to encourage us to engage in open ended, introspective, and goalless chats.

Where did this idea come from? Well, in case it wasn't obvious, it came from the likes of Chai and Character AI and Replica. Imagine Sam's shock when he learned that people were spending 90 minutes a day on average having aimless conversations with lobotomized Llama models and paying large amounts of money for the privilege. Not everyone is on board with this pivot from OpenAI. Reactions are split, probably more negative. In some quarters, the scorn is almost as rancid as when Facebook famously moved the newsfeed to be based on recommendations or collaborative filtering away from chronological. Some users love the human touch. Others prefer Chatchy BT's older utilitarian vibes. The sycophancy and memory is the devil incarnated for any technical query because all this information serves as a distractor and demonstrably deteriorates the results. There's also the small matter of personality versus intelligence. Clearly, they are inversely proportional to each other. How many brilliant people do you know without a charisma bypass? Remember that book, how to win friends and influence people by Dale Carnegie? It showed in excruciating detail that flattery gets you everywhere in life, and you should never get into arguments to get ahead in life. And let's not forget the famous Eliza chatbot, the grandmother of all chatbots developed in the 19 sixties. It used very simple pattern matching and keyword substitution to mimic a psychotherapist. It had 0 genuine understanding, yet people confided in it, felt heard by it, and even became emotionally attached. Joseph Weisenbaum, its creator, was famously horrified by this reaction. He saw how easily humans projected genuine understanding and empathy onto simplified algorithmic tricks. Fast forward 60 years and we have vastly more sophisticated pattern matching language models. These days, the ELISA effect can be an extremely profitable tool for charm and simulated empathy for maximum engagement, but it's also more than that. It has the genuine potential to improve lives. Now the audience of MLST, we're probably a little bit ahead of the curve on LLMs compared to the average person. I mean, God knows we've been using LLMs day in, day out. I have for the last 5 years. And we expect them to be correct, not sycophantic. The other stark departure with FourO is that like Chai, they are optimized based on engagement over time, which is to say the average conversation length, or at least that's what I assume they are doing. I don't know for sure. Before, it was just RLHF on isolated conversation threads. This is an important first step towards AI becoming a social network. The Harvard Business Review recently showed that AI companionship and therapy was the number 1 use case for artificial intelligence in 2025. And of course, you might remember the interview we did with Daniel Khan at Slingshot AI. They're building a a therapy bot last summer.

Daniel Cahn 00:47:29

There is a real category called therapy that actually helps people that can be more effective than drugs. Like, a fully optimistic, self determination esque people who are thinking about their own agency are choosing to go to therapy, and it actually does help them.

Tim Scarfe 00:47:43

Yeah, it does. But in in 20 percent of cases, it makes it worse.

Daniel Cahn 00:47:46

Point of the negative is to say that if you have real medicine, if something really does work, it also has risks. Obviously, you cut open a person's body to remove a tumor, you can also kill them because you opened up their body.

Clip Speaker 00:47:55

I would basically just guess that you go forward a few years, we're just gonna be talking to

The Future of AI Intimacy

Clip Speaker 00:48:00

AI throughout the day about different things that we're wondering. It's like you'll have your phone. You'll talk to it on your phone. You'll talk to it while you're browsing your feed apps. It'll give you context about different stuff. You'll be able to answer questions. Help you as you're interacting with people in messaging apps. O3 has a very different skill set. It can think through problems really hard. You don't really want the model to think for 5 minutes when you say hi. And so I think the real challenge facing us on post training and research more broadly is, like, combining these capabilities. So, you know, training the model to be, like, just a really delightful chitchat partner, but also know when to, you know, reason.

William Beauchamp 00:48:45

Talking about Sam Altman a little bit. You know he's building all of these social features in and kinda taking a step in in this direction. Are are you worried about that? I think competition is fantastic. I think competition forces everyone to step their game up, whether that's DeepSeek releasing their really cool research and their model gave everyone, you know, on this side of the Pacific a

little wake up call, and I think everyone has kind of stepped their game up a bit. It's the same at Chide. You know, there's been waves of competition. You know, we've definitely enjoyed periods of being number 1. Right? And then someone's come along and shown us 1 or 2 things, and then we found ourselves in that uncomfortable uncomfortable position of being number 2. And it keeps the team hungry. It keeps you searching for new ideas, pushing the boundaries, pushing the frontier. So I'm a big fan of competition. I think it's great. I think a lot of people forget, like, how big AI is and how big the space is. And I think you can look at video and you can say there's a YouTube, there's a TikTok, there's a Netflix, there's Amazon Prime, there's Disney, there's Apple. I think that's what AI will look like. I think there will be so many multi 100,000,000,000, trillion dollar businesses that are being created. So I wouldn't I you know, we don't get scared of competition,

Tim Scarfe 00:50:05

but instead we wanna challenge ourselves to be ahead of it. We started by asking what happens when the line between human connection and artificial intimacy blurs. Chai's journey gave us the early answers. Millions flocked to AI simulators, imagination machines seeking laughter, solace, companionship, and occasionally, sexting. So the genie's out of the bottle. The message is unavoidable. AI is no longer just about answers. It's about us. It's about connections, simulated or otherwise. Chai proved the hunger was there, building a fiercely loyal user base by, as Will Beauchamp put it, trusting the community over elites. And now OpenAI's pivot with GPT 4 0 means the space is only getting larger.

References

[1] **Black Mirror: Be Right Back (S2E1)**
https://en.wikipedia.org/wiki/Be_Right_Back

[2] **Tufa AI Labs**
https://tufalabs.ai/

[3] **AI Chatbots for Depression and Anxiety Meta-analysis**
https://pubmed.ncbi.nlm.nih.gov/39162424/

[4] **Woebot Health**
https://woebothealth.com/

[5] **Sam Altman TED Talk with Chris Anderson**
https://www.youtube.com/watch?v=6Kp_yxwnVCk

[6] **Chai Research - Careers**
https://www.chai-research.com/jobs/