**Aravind:** 0:00

We just built it as a tool that was useful to us because we kind of had no idea of how to build a company, how to build a product. Neither three of us had done anything of this nature, and we initially built the tool like chat GPT, where it'd be like a slack bot and then we would be able to ask it a bunch of questions and it would just give you the answer. We realised that it would hallucinate a lot, it would just make up stuff and it was not trustworthy. And my co-founder had this idea that, ok, what if you ground it in real web links? And then we tested early feedback and then the people who used it in the beginning, they were all like, hey, look, I know that you just built it as a cool tool, but I actually find this to be better than Google.

**Craig Smith:** 0:43

Hi, I'm Craig Smith and this is Eye on AI. In this episode, I talk with Arvind Srinivas, co-founder and CEO of Perplexity AI. Perplexity is reshaping the search landscape with what it calls its conversational answer engine, providing direct answers grounded in real facts. The company has made a splash by claiming its goal is to supplant Google, but Arvind is realistic about the challenges. He talks about building a reliable search and ranking system to ground Perplexity's large language model in trusted information. I hope you find the conversation as fascinating as I did.

**Aravind:** 1:28

I'm Arvind Srinivas. I'm the co-founder and CEO of Perplexity. Perplexity is a conversational answer engine that directly seeks to give you an answer to any question that you have, and the way it differentiates from existing tools like Google, like search engines, is that you get the answer instead of getting a bunch of links and having to open them and read them yourself. And the way it differentiates itself from other chat tools like chat GPT that also directly answer your question is that it's grounded in real facts and tells the user exactly where it's pulling its content in the answer from, in the form of citations or references. So it's like if Wikipedia and chat GPT had a baby and used all of the data from the internet, that product would look exactly like Perplexity, and that's what we are.

 **Craig Smith:** 2:20

Yeah, and how did you come to this? What's your educational and career trajectory?

 **Aravind:** 2:26

Yeah, so I originally come from India. I was born in India, I lived there for the first 22 years of my life and I came to the United States in 2017. I had gotten undergrad in electrical engineering with a lot of focus on computer science and machine learning from Indian Institute of Technology Madras IIT, madras and it's one of the top engineering colleges there in India, and during my undergrad period I got to intern in many American universities and Canadian universities. One of them was Turing Award winner lab, Joshua Bengio. He's one of the deep learning pioneers and I got really excited by that and I wanted to do a PhD combining deep learning and reinforcement learning, because deep mind was the hottest lab in town at the time and they were ruling the rules with all their real progress. So I really wanted to do more research there and I got admitted to Berkeley, which is again one of the top American universities, particularly for AI, and OpenAI had a strong connection to Berkeley. They used to recruit a lot from Berkeley, so I got lucky there and got to intern at OpenAI at the end of my first year and that was when I got to meet amazing, amazing scientists and researchers there. Folks like Ilya Sutskever were chief scientist and then Alec Radford, and they were building GPT-1.

**Aravind:** 4:17

At the time it wasn't called GPT-1. It was just called GPT. Nobody knew there would be so many generations of it and it was clear it was something very different. So I really wanted to study more about it. Try to change my topic in research from RL to generative models, generative AI. We used to call it unsupervised learning at the time, so I kind of like trying to do more of what we were thinking, and I got to do a pretty good PhD. We published a lot of good papers on transformers and the highlight of my PhD was working with people at DeepMind during my PhD and DeepMind now Google Brain, but now it's all the same and during one of my internships at DeepMind in London I used to be launching jobs during the day and then I used to be in the library during the evening because as an intern, you just want to stay in the office as much as you can and learn as much as possible. So I did that and I stumbled upon this book called how Google Works by Eric Schmidt and Indeplex and how Google Works. These were two different books and I really loved the entrepreneurship conversion from research.

**Aravind:** 5:52

I was always excited about entrepreneurship. I watched movies like Pirates in Silicon Valley and TV shows like Silicon Valley, but never thought I was one of those people, because all the examples there like Mark Zuckerberg or Bill Gates or Steve Jobs, were all like undergraduate dropouts and I was here, already completed my undergrad, right like I could only get to the US because I completed my undergrad. And so I was like, OK, you know what it's over? I'm already. I'm only born to work for these people. I'm not born to actually do a company myself.

**Aravind:** 6:30

But then I looked at this example of Larry Page and Sergey Brin and they were basically PhD students, completed their undergrad and then got an idea and, while working on their PhD, and made it into a legendary company where I'm actually interning in at that point. So that really appealed to me and I went and asked one of my mentors at Google hey, like, what do you think is the page rank of you know 2019? And you know we would talk about what is the most important idea and everybody would just say transformers. And this was still at a time when there was GPT-2 and there were still a lot of sceptics at DeepMind. But like, oh what? Like it's cool, but it's not really intelligent. It's just, like you know, autocomplete, but again, like, look how far we've come from there. So I went and reached out to the inventor of transformers for an internship. He was a Google brain.

**Craig Smith:** 7:30

Yeah, who is that that you reach at Lucas?

**Aravind:** 7:33

This is the first author, Ashish Vaswani Like, look at this, he's a co-contributor. I got to work with him and learn so much from him that, like you know, super grateful, he's also an investor in perplexity. Yeah, so we worked on making transformers like a universal computation system, not just for natural language but also for computer vision, for our realm. And I was very inspired by this way of thinking about how like transformers are, like you know, one universal architecture for everything. And that got me into whether I was deciding between at the end of my internship my PhD was also coming to an end, so I was deciding between either becoming a research scientist or starting a company, because I always wanted to start one. And now that I understand transformers pretty well, it's time to go and build a product using it. But the conviction was not there. Like Transformers was still sort of something where you would see research progress but nothing like a product like GPT, you know, like 3.5 or you know Dall-E, none of these things existed at the time. There was a GPT-3, it was a toy. You could see the kind of work, but most of the time it wouldn't work. So that conviction was not there in anybody to truly go and launch a product and there was no example of any successful business being built with it either. So I decided I would just stick with research. I also had a lot of immigration issues, visa issues, so I just stuck with research. Now I joined OpenAI after my PhD to work more on, like to learn more about another class of models. That was also very important in general, the way I called diffusion models. So that was what I was spending time on as a researcher. But then I got to see, you know, like companies like Jasper and CopyAI and GitHub, copilot all emerged as very successful products being built on top of GPT-3.5 and making a lot of revenue like a lot more revenue than even OpenAI was making itself.

**Aravind:** 9:44

And that moment made it very clear to me okay, look, your entrepreneurship time has arrived. I was already like 27, 28 at the time and I was like look, I don't have too many more years of high intensity work left in my body. Like I can work super hard even in my 30s. Like I can work really, really hard, but I don't have the 100 hour work week, week after week after week, for like five years of grind left. My mind won't work, even if my body wants it. My mind will get burnt out. I'll get a zone. I'll zone out at meetings. I won't be able to digest complex concepts as fast anymore. I won't be able to code for many hours together anymore. So I realised that and felt like this was now or never for me.

**Aravind:** 10:41

And also by mid 30s you usually kind of build a family stuff like that, right? Like there are so many other constraints outside of yourself, and so I took the plunge. Okay, the market is there, products are being built, technology is improving, costs are going to come down, the models are only going to get better. So now is the time. So I decided to leave and I called a few investors like Elad Gil and Nat Friedman. Elad Gil is a legendary Silicon Valley investor, basically every unicorn in the internet and mobile era. You could look at it and he would be an investor in an Airbnb Stripe, dropbox, OpenDoor I don't know Pinterest, figma he's an investor in all of these. And so, like I emailed him on LinkedIn and he just responded in an R and then he converted me into a founder very fast, like he's very good at convincing people to leave. And Nat Friedman too.

**Craig Smith:** 11:48

I'm just curious what the LinkedIn message was?

**Aravind:** 11:53

I don't know. I can't remember this specific wording, but it was more like hey, elad, I'm working at OpenAI, I work on all these diffusion models and stuff. I'm a researcher but I want to actually start a company. That's what my heart is in, and I don't have any idea for any product or market. I have, like, literally zero clue. It's not my world. I'm coming from an academic world, but I would love to use your advice on figuring it out. If you can spare half an hour, that would be great.

**Aravind:** 12:33

And he replied very instantly, you know. So that actually sometimes I'm very grateful, like, why would? Why do people care about some new person you've never heard about? Of course, I had the track record and all that stuff, so that definitely helps, but still, someone has to take their time and do these things and I deeply I'm grateful for him for that. And then Nat Friedman too. I Twitter DMed him and he just responded very fast and the fact that Elad was already talking to him definitely helped. And so we, these two, committed to invest in me without any idea. Like I just had a few demos and they said okay, look, we can put a million dollars together, like one to 1.5 million or something of that nature, and friends past friends also were supportive and they together we pulled together like two to 2.5 million and started the company. Now you can ask like, is that a good amount of funding? It's very little funding, honestly, compared to other AI companies that don't have anything and just go and raise tens of hundreds of millions directly.

**Craig Smith:** 13:51

When was this initial aim?

**Aravind:** 13:53

June, july. We started talking around June, July and the funding came together in September.

**Craig Smith:** 14:02

What year is that? 2022. Before chatGPT.

**Aravind:** 14:05

Now you can say, if I knew chatGPT was going to happen, it would have been a dumb decision to leave OpenAI. Right, you never knew. Like that. That's just kind of how the world works. You make your own luck and then you can ask, do I regret it? Would I rather be in an OpenAI than to do my own company? I don't regret it at all. It's just, in fact, I feel more proud of myself that, like, despite such a massive success, perplexity itself has its own name and like a different value proposition in the field, in the product space, and you truly want to build your muscle and prove to yourself you're capable of a bigger challenge by doing your own company. The amount of energy and commitment it takes to do this is just a lot more than writing a small component of a large system. That gets the company, like another larger company, a lot of credit. Yeah, so that's kind of how it all came together.

**Aravind:** 15:11

I knew my co-founder, Dennis Yaratz, from my PhD days in Berkeley. We wrote the same paper but, like independently, we discovered the same thing and wrote the same paper one or two days apart. And then he brought in another co-founder, who I did not know before but he knew it's just Johnny Ho, who was like world number one at competitive programming and they had worked together in this company called Cora, which was a question answering site. So that's kind of how it all came together for us and to date, like you know, I think we all three have very complementary skills, which is pretty rare in a founding team. Usually you're trying to start with people you know and then try to scope out roles. In our case there was at least a pair of people who knew each other, but it's not like we three had worked together. In fact, Dennis and Johnny themselves hadn't talked to each other for a while because they were like not working together for a while.

**Aravind:** 16:15

But they have different skills. Like Dennis is very good at prototyping and testing new ideas quickly, Johnny is very good at actually building systems back end and strategizing and likes being very methodical about things. I am very good at big picture and vision and like actually convincing people, recruiting, fundraising and, like you know, positioning the product well, talking to users, like those kinds of things. So it was like the perfect team, in my opinion, to do a company in space. And all three of us are very fast iterators, like we don't like stasis. We just kind of like one constant in iteration and improvement week. So that helped you kind of feed off each other's energy in the beginning and that really helped a lot. And, yeah, we stuck through. There were so many challenges, so many difficult moments, so many times when I felt like why am I even doing this? But you know, it's all worth it to make the planet smarter.

**Craig Smith:** 17:34

Sure, and so you've raised over half a billion at this point. But let me ask you about the product. So this is Perplexity AI, and you described it as a question-answer platform. What I don't understand about Perplexity I understand a little bit about what you're doing in the background, but let me ask you first who came up with the message that you're a Google competitor, you're taking on Google.

**Aravind:** 18:10

So I have never really wanted to do that, to be very honest.

**Craig Smith:** 18:17

It's a pretty bold statement. Yeah, that's one of the reasons I said at the beginning I want to talk to you about your PR. I mean, where did that messaging come from? Because, yeah, it gets a lot of attention, certainly.

**Aravind:** 18:32

I think I would be telling the truth here, so, okay. So I guess the truth is that when we built this tool initially, we just built it as a tool that was useful to us, because we kind of had no idea of how to build a company, how to build a product. Not, neither three of us had done anything of this nature, and we initially built the tool like chat GPT, where it'd be like a slack bot and then we would be able to ask it a bunch of questions and it would just give you the answer by opening a new thread in response to the question you ask. And we realised that it would hallucinate a lot. It would just make up stuff. It was not trustworthy and my co-founder had this idea that, okay, like what if you ground it in real web links? They said, okay, we both are academics. What do academics do? What do journalists do? And they write a paper or write a story. They only write stuff for which they actually have a reference. You're not allowed to say and make up stuff, right, that's the ethics in journalism and that's the ethics in academia. What if we bring that personality into an AI chatbot like a chat GPT that always looks up sources.

**Aravind:** 19:53

That was the idea and that ended up becoming a very useful bot of its own, that we made it a discord bot with a new discord server, and then we tested early feedback and then the people who used it in the beginning, they were all like hey look, I know that you just built it as a cool tool, but I actually find this to be better than Google. The users said it. So when we released a real like hey look, guys, we made this cool new tool. That's like a GPT chat. That's always accurate because it uses real links and facts and you get citations. That's how we released it. But then people were like hey look, this is not really a chat GPT competitor, because chat GPT is very useful on its own, even if it doesn't use real links. It's very useful because it's a brainstorming buddy.

**Aravind:** 20:46

So I still want to be just using it for generating content, where hallucinations was a feature, not a bug, whereas your product is small, like hallucinations is a bug. You want it to be as accurate as possible and that's why you're using real links and you're always retrieving and then generating. So I feel like you're more of a Google competitor because you're trying to save time. You're trying to save people time from reading a bunch of web pages and like trying to distil the content and those web pages on your own. Instead, the AI does their work for you now, and then there's no need to type in keywords anymore, you just directly get to the point. So you're actually doing a paradigm shift to Google. So that's how we realised it.

**Aravind:** 21:30

After rolling out the product, we realised that we have something actually much bigger than what we thought. And then further reasoning this is something we never had any idea about in the beginning which is, hey, it's not just a better product experience over Google, it actually disrupts Google's business model. Like, you can have something better than Google on search and get killed right away, right, if Google just rolls out the same thing. Example let's say you roll out a feature where, let's say, Google hasn't shipped speech to speech experiences and you roll out a feature that literally is Google plus voice to voice, you can get killed. That's fine. But if you fundamentally say I don't need links anymore, I just you know, I'm just going to give you the answer directly then that disrupts their display advertising business model, because they basically tied their business model with their front end, with their UI, the 10 blue links. Which link gets shown at the top? How many people view a link? How many people click on a link? This is analytics that's fed to an advertiser. So, and when I'm telling you, hey, that UI is going to go away, and now you're going to see something else, and answer is the prominent part of the UI. Now that is a very advantageous position to be in, because you are playing to their weakness, which is the deep tie with this UI, on which you're making a trillion dollar market cap. Like, if you remember, two weeks ago, they went, they, they, they. Advertising revenue increased, but it was a billion dollar short of Wall Street expectations, which expected them to increase even more. And even though they made up for falling short of the expectations by making up revenue in subscription and cloud, the stock price went down by six percent. Now that is the, that is their problem, that is the, that is their problem and that's why we are able to operate.

**Aravind:** 23:44

This innovator dilemma is playing out.

**Aravind:** 23:47

And then we decided okay, look, we stumbled upon a magical idea.

**Aravind:** 23:52

Everybody wanted to take on Google for the last 20 years.

**Aravind:** 23:56

Nobody succeeded because there was no technology that enabled you to change the UI of search and now, finally, it's getting changed, and what's more magical is that technology is in better shape outside Google than internally inside Google for the first time Ever, ever, ever since the last 10 years, like for every single time the last 10 years, the best AI models always lived inside Google. So, even if there was a way to disrupt the Google experience, only Google could do it and therefore they were safe. But because of open AI, because of all these open source AI models, suddenly, like the rest of the world is leapfrog, Google and AI, and therefore a startup like us, can actually boldly try to attempt to Bring in this disruption. And so we, we got very lucky to be in the right place at the right time, right after chat, gpt, gpt, 3.5, lama, to all these things are happening and Also, at the same time, google was going through bureaucracy and, like political problems, slow progress, and so we were like, so lucky to just, you know, take advantage of the time.

**Craig Smith:** 25:11

You're searching a curated set of Sources. That's how you ensure precision or correctness. Bing already does the same thing. They search, they use Bing search, they list the links, just like you guys do. And it's you know. You set it to be more precise and it's, it's pretty precise. So you guys hit a moment when you were able to raise money. But how do you survive in competing Against Bing, in competing against Gemini, which is most certainly going to do the same thing? How do you deal with all that?

**Aravind:** 25:55

All right, yeah, let me. Let me go one by one, right? So the first, that first point about Bing. Then there was a point when I was very scared of Bing? Which was on the day I signed the term sheet for a series a? Literally we had a handshake with our investor and we were chilling in a cafe in Palo Alto Because that week was so stressful. And we see the words Leaking, screenshots of Bing, the new version of Bing, and we are like, okay, that's it. The VC is going to go back on the word. And you know, there's a 30 day due diligence period where the VCs check the market and make a decision on whether to wire the money. And they're not going to do it, and I would not, you know, I just have to go and hope Microsoft or Google buys off for like a little bit of money so that I at least save my face here. That was the extent of fear at that time.

**Aravind:** 26:59

But Bing bought, they tried to do a free form chatbot and a search bot all in one, one single product. They did not explicitly focus on the search use case. They rolled out. They had the advantage of having access to the most cutting edge model at the time, gpt-4, whereas everybody else in the world had only GPT-3.5. Despite that, despite all those advantages, all the distribution advantages, marketing, or they bottled it because of poor product execution and the fact that Bing has always been associated with a bad consumer product. The brand of Bing has always been associated with something that was like second tier. The third reason that they're not taking office is that they want users to install the edge browser or have a Microsoft account to be able to use Bing chat. So all these things stopped them from taking the first more advantage and gave the opportunity for people like us to still exist. And outside of that, people view perplexity as a much faster and better done consumer product than Bing. Our mobile apps are much more natively done on SwiftUI, whereas Bing apps are built on Flutter. So the apps are much bigger in terms of memory consumption and size, slower, less well executed.

**Aravind:** 28:27

Their focus is actually very different. Their focus is not really on delivering the best experience. Their focus is just getting browser market share. It's a very Microsoft way of going about things. Hey, I just want to get edge users. I just want to get more edge users than Chrome. So what am I going to do? I'm going to offer people who are edge users a lot of freebies, like GPT-4, and then, like Dali, or like GPT-4 vision or, like you know, pdf summaries on the side, all these kinds of things. That way, I'm hoping that all my AI ownership of OpenAI translates to browser market share, which might translate to advertising revenue increase.

**Aravind:** 29:08

So it's a very Microsoft way of thinking about it, whereas, honestly, what is the real thing you should do? Deliver the best possible answer engine, the best end-to-end product experience to the user. Now, they don't have any incentive to do that because they are a trillion dollar company that's only trying to make a difference in terms of tens of billions of dollars in revenue. For us no revenue. Small startups focus on product nail it, get the user base incrementally right. So that's why, like we did much better than Bing and we could, you know, to add fuel to my arguments, they even changed Bing chat to call it co-pilot. They realised it didn't work and now they're trying to adopt a branding of some word Microsoft co-pilot, which is typically associated with an enterprise. Saas product Not doesn't work for a consumer, so it didn't really work for them and our bet ended up being right.

**Aravind:** 30:08

Now, as for Gemini, I guess what I'll say is like Google is making the same mistakes they've been making for a while and messaging apps. They're constantly changing and rebranding things and having multiple competing products existing at the same time. There is Google search generative experience, sge that is on Google.com itself and arbitrarily fires and doesn't fire based on whether you turn it off or not and doesn't fire for any commercial intent queries actually. And then there's Google Barred, which is now rebranding to Gemini, and there's AI that's also trying to do something like that for enterprises. So they all also like different versions of the same product that are being rolled out.

**Aravind:** 30:59

And when people think about Bard, the branding that Google has gone for is like your thought partner, your buddy or something like that, more as a competition for chat GPT trying to do one single chatbot for all purposes freeform a search doesn't matter, but they're trying to promote it more for freeform chat brainstorming, whereas perplexity is just purely focused on search, like I'm not even trying to be a chatbot alternative, I'm trying to be the most accurate chatbot ever, like most accurate answer bot. So when people think about perplexity, think about accuracy and fast, reliable answers. They think about so many things at once. So that is a key difference.

**Aravind:** 31:43

And the other thing I would say is Google has no incentive. Like, honestly, look, why would? If you are Google, why would you build user base from zero for a new product? Like that's basically you not making use of all your distribution advantages. Right, if you were, if you were Google, you would just directly roll out to the billions that you've already accumulated over two decades. But Bard built a user base from zero.

**Aravind:** 32:12

Of course, they had the benefit of Google marketing and PR, but it's like a very suboptimal move for a big tech company to do that. And why did they do that? Because they cannot do the other thing. They cannot do the other thing of changing Google.com to look like Bard. Okay, if Google.com looked exactly like Bard today, perplexity is dead, we're done. But Google is also dead. We are building our own search. I mean, we've been building our own search ever since the beginning. We rely on some ranking signals that we can scrape from Google search results pages, but we're not overly dependent on them in any way that our service wouldn't be able to run without them or something.

**Craig Smith:** 32:51

When you fire a query, the back end doesn't send it to Google.

**Aravind:** 32:56

So Google has no API you can scrape search result pages of Google. There's no official API for Google Now and if you only scrape search result pages of Google, your latency would be very bad, it'll be very slow and you have the same problems that Google has in terms of SEO, spam and ads and things like that. So the fact that we don't suffer from this problem clearly indicates that we are doing a lot more than just hitting Google API, which, of course, doesn't exist. But even if you consider hitting Google API, meaning just getting whatever the search results pages offer, clearly we're doing a lot more than that and we have been building our own infrastructure for picking results from our own pages. And if you just hit Google API, you're bottlenecked by what the snippets are given to you from the search results page, which are going to be low quality and less informative and will lead to a lot of hallucinations.

**Aravind:** 33:53

So you have to build your own index, you have to go and crawl the web yourself and you have to build your own ranking and you have to build your own page rank of the web for which domains to use and not use, and we've been doing this ever since the beginning. Now this is a multi-year journey. I'm not telling you we're done. I'm not giving you any indication that we have something that rivals Google. All I'm telling you is that this is going to take a while, and the only way to win here is to have a product that's already being used by people on a day to day basis and use all that data flywheel to build an amazing ranking and index, which will end up being like technological modes for you.

**Craig Smith:** 34:37

And so on, this crawling the web and indexing sites for use by perplexity. That's got to be automated. What metrics are you using to rank authority or accuracy? Because that's what it comes down to is having a curated list of sites.

**Aravind:** 35:10

I would say that initially we started off doing a lot of manual work for this, like I literally wrote down some domains that we just had to not use and use and stuff like that.

**Aravind:** 35:22

But obviously the process is not scalable and we have a more sophisticated process that uses our own aggregate data in terms of what domains get linked at the top as citations.

**Aravind:** 35:40

You can think about what domains get cited as a different signal for page rank of the web, and then the frequency of citations can decide how often you should prioritise using a domain's results versus another domain, and then, honestly, that alone is not enough. You have to use a ton of other signals in terms of prioritising the ranking, including your confidence of how good the snippet from the domain will be, based on your own crawling infrastructure. So we are constantly upgrading and updating these systems as we speak. We have a very, very, very talented search and ranking team that's working all the time here, and I believe whatever solution we'll implement will constantly need to get upgraded, because the web keeps changing every day and there are different ways to fall into the trap of poorer quality results. That's why this company is hard. Sometimes I think why should this even be a company versus just being a project. The difference between a project and a company is a project gets done, there is a deliverable and it's over.

**Craig Smith:** 36:53

This process of indexing Perplexity when I use it every now and then I'll ask it something and it doesn't have an answer, presumably because the pages index don't include that, and then I switch frankly to Bing Co-Pilot when you switch to Bing Co-Pilot.

**Aravind:** 37:13

Do you get the answer?

**Craig Smith:** 37:14

Usually yeah, there are other times when it's the reverse, when I have to ask Bing twice to get the answer. I won't find it the first time. But I mean, I don't know how you measure it, but how much have you indexed based on how much you want to index? Are you 10% there? Are you 2% there?

**Aravind:** 37:45

50% there, I would say we are 70% to 80% there. In terms of really important pages on the web, we are able to prioritise the citations. Basically, they are OK here. Maybe I should add a nuance here. I don't want to say we have 70% of the whole web's index captured. What I want to say is there is a good part of the web. A small portion of the web is actually useful. Now in that portion we have captured 70% to 80% of the value already. There is a long tail. Even there Now the web is incredibly long tail. You only want the head of the distribution. Even within the head of the distribution there is a long tail. So that's what we're going to try to capture, and if we have captured the basic head of the whole web, I feel like it's a game set match for us in the next generation of the way people are consuming information, because nobody wants the whole web anymore.

**Craig Smith:** 38:48

And then another question. My career was with the New York Times and there's this lawsuit with OpenAI. How do you feel about that? Because if you're indexing and pulling data from sites, personally I think fair use will win out. But how do you feel about that? Because the New York Times argues that you're stealing the not you, I mean that OpenAI is stealing the revenue from New York Times because the users see the text response. They don't see the links. Or if they do see the links, they don't see the advertising on the links.

**Aravind:** 39:47

Yeah, I would say that OpenAI steals other people's revenue a lot more is an argument that you can empathise with, because they don't actually display the sources and they've trained all these models and all the data and, leaving aside the training, even the end product the chatGPT end product doesn't display the sources. So then, if it literally reproduces the content that was there in the original publisher's site without actually attributing it to them, I do feel like that's something that should be taken seriously. So the New York Times has some valid arguments there. Now, in our case, we attribute the sources directly, so we're not actually saying this content is from perplexity. It's actually just the users, the publishers, content that we are just making sure it reaches more users when they're asking something else. So the content publisher should feel good about it and we can still drive traffic to them, and perplexity does drive traffic to others. Have you signed a contract with the publisher?

**Craig Smith:** 41:03

Have you signed any agreements with any of the sites that you index.

**Aravind:** 41:12

We don't have licensing agreements, so we have informal equity partnerships with anybody. But we do pay some data providers for API access and I anticipate doing a lot more of that nature.

**Craig Smith:** 41:29

Did the idea of building a vector database ever occur to you guys as not obviously as the main source, but as an auxiliary knowledge base?

**Aravind:** 41:43

A vector database can be constructed as a proprietary vector embedding that you have, and then the algorithms for fast retrieval, fast key value retrieval. Now the second part doesn't need anything proprietary from your end. Your people outside are doing it really well, and there we use this thing called Quadrant, which is really fast and very good, but the embeddings that are used for Quadrant are trained by ourselves, and so in some sense, you can see, we have built a vector database, the end to end vector database, but we rely on the infrastructure that's being done by other people, and Twitter X is also using Quadrant, I believe.

**Craig Smith:** 42:21

Yeah, and the things that the embeddings in the for you in Quadrant are the URLs of trusted sites or they're the content?

**Aravind:** 42:33

Yeah, so the traditional key value database is URLs and you don't need a vector database for that. Now the actual content itself is chunked into pieces and that requires your vector database. Vector embeddings.

**Craig Smith:** 42:48

Yeah, that's fascinating, and so you're. You're, you're aware of the looming threats, but you think you have an opportunity to scoot through. You've got plenty of funding. I was interested that the basers invested, because Amazon's invested in Anthropic and again, it seems as an anthropic, together with Amazon, could, could steamroll you guys. But how do you feel about that? Anthropic is still kind of trying to find its market fit.

**Aravind:** 43:29

So I would disagree with your conclusion that Anthropic can steamroll us with Amazon because they don't have any expertise in search technology. They're only working on chat bot and then clawed the model that's powering the chat bot and even leaving aside search, their work so far and building a first party consumer product has been like not that great compared to even players like us. So I'm not really worried about Anthropic and also Jeff's investment in us has nothing to do with Amazon's investment in Anthropic.

**Craig Smith:** 44:03

And so what's next? You can continue indexing.

**Aravind:** 44:06

Yeah, it's very simple, actually. A lot of people ask me this question: what is next? What is? You know, what's one of the two key things that will change your destiny? And I give them the very boring answer: we just need to keep growing. And the best way to grow is to make the product better. And what does it mean to make the product better is to make it faster, more accurate and make the answers more readable. So faster, more accurate, more readable, more presentable, and if we just deliver on this, then our growth and retention will automatically improve and that will lead, to continue to lead to, like you know, strong word of mouth growth, as it's like marketing opportunities, and that will further drive home more awareness. And then that will make all the underlying models and indexes better, bigger and more useful, and this will feed, this will continue to feed each other the work we do on making the product better and the work we do on making the index better. So they will feed each other.

**Aravind:** 45:15

And then, if we can create that ever self fulfilling cycle, flywheel and keep hiring well, I feel like we have a real shot at success. So that's my job and some machines keep running.

**Craig Smith:** 45:34

Yeah, that is. Do you have much of an API business? Is that revenue?

**Aravind:** 45:42

Yeah, we have our own proprietary PPLX API that has a first of its kind online LLMs without any knowledge cut off, so it always retrieves from our index and in rights the answer, and I believe that API will see a lot of adoption as we increase the rate limits.

**Aravind:** 46:05

Currently, the rate limits are very strict because I would say, like reduce the rate limits, like you know, basically make it possible for more people to submit more requests per second or minute, because right now it's severely rate limited because of our own infrastructure. We don't have many GPUs, so we, now that we have more funding, you can actually expand access and then, once we do that, I believe, like our API, business will also flourish and I think that's very important so that it doesn't put us in a pressure to figure out like appetising revenue soon. Imagine if Google had Google Cloud figured out before Amazon figured out AWS. They wouldn't have needed to go this deep into advertising. So, considering the hindsight of how the world is played out, you know, I think we have the luxury of trying many different business models at once.

**Craig Smith:** 47:09

Yeah, the. What kind of companies are using the API at this point?

**Aravind:** 47:15

I do not have enterprise contracts signed with big companies yet. I would say we have around 10,000 developers that are building with our APIs, spread across so many different companies, and I think we'll have to mature more as a company to do large enterprise contracts with, like established companies.

**Craig Smith:** 47:35

Yeah, you know, at the beginning you were saying and I've read that you're building on GPT 3.5. Now you have GPT 4 in your paid version, I guess. But then you have a LLaMA 2, or you Fine-tuned the model off of a LLaMA. What is LLaMA? What role does LLaMA have? Does the LLaMA 2 model play?

**Aravind:** 48:03

Yeah, so we build on top of two open-source models LLaMA 2, and then this straw, mistral and mixed straws. A new company called mistral that's published a bunch of cool models and the role that these models play is to give startups like us a chance to build on a stack that has no Big Companies influence. So, for example, open AI we are competing with them on, you know, consumer products like chatGPT and for flexi, about trying to, you know, compete for the same user subscription revenue. So Having alternatives to model access outside open AI is very important to us, so that's a role being played by Anthropics cloud or this is a closed source model, or open source models like llama and mistral, the.

**Aravind:** 48:56

The additional advantage of open source, obviously, is the fact that you have access to the weights and so you have the flexibility to change these models even more To suit your product and further differentiate your product so that it doesn't exactly respond in the same style and idiosyncrasies of GPT 3.5 and4, doesn't moralise people has, is a lot less restricted and therefore Can, like you know, attract a different user base or the same user base even that might want like different kind of behaviour. So it's very good. It's very good for the world that, like meta and miss Rall, like, focused on open sourcing models.

**Craig Smith:** 49:39

Yeah, yeah, no, I agree. And so these open source models, llama and mistral, are playing the same role as GPT 3.5. It's just: Do you split a query when it comes in, and send some of it to the open source and some of it to GPT 3.5? And then, you know, put them back together when, when you get the response, or do you, depending on rate levels or traffic to use?

**Aravind:** 50:11

Yeah, yeah, we are our router infrastructure. It's so sophisticated today that I Don't even know what we're currently doing. It's something that keeps intelligently changing based on the throughput and the latency, and like Whether some existing API endpoint is down or up. And we we do like a lot of intelligent routing, but our goal are, like maybe I'll give you a more broader goal is to be able to run like if, if, in a very Near future term, like we should be able to achieve this goal that our model is the best model for our product, and the singular reason for that is because we have so much user data that we should be able to specialise our model for our product and it should be a lot better than making use of open as models for that purpose or or anthropics model. And we should also be able to achieve like true independence that we like oh you know, it's not just for saying you're independent, but it's actually in your interest to be because you, you cannot build this product in the most useful way to your user without actually tuning it to them and therefore you're achieved something truly remarkable. And that requires you to not just be able to train models. Well, you should also be able to serve them.

**Aravind:** 51:36

Inference infrastructure should be very robust and scalable, which we've achieved already and now we'll soon achieve the training milestones too and then, like with the advent of better open source models like llama 3 and future versions of mistral, we should be able to Achieve the parity with not just GPT 3.5, but also with four, and that'll be remarkable because then in a in that world then you would say Open AI or anthropic should have even better models in GPT 4 and the. What kind of search experiences can you create with those models that you know you've not seen before? So I always feel like that'll end up being how the world works is like Whatever is possible today will be possible with our own models very fast, but then that'd be a more intelligent model out there in the market and then we will also innovate on the product part and see what kind of search experiences that were never possible before that can certainly be possible.

**Craig Smith:** 52:35

That's it for this episode. I want to thank Arvind for his time. If you want to read a transcript of today's conversation. You can find one on our website. Eye on AI, that's EYE-ON ai. And in the meantime, remember the singularity may not be near, but AI is changing our world, so pay attention.