Welcome to the OpenAI podcast, the podcast that opens up the world of AI in a quick and concise manner.
Tune in daily to hear the latest news and breakthroughs in the rapidly evolving world of artificial intelligence.
If you've been following the podcast for a while, you'll know that over the last six months I've been working on a stealth AI startup.
Of the hundreds of projects I've covered, this is the one that I believe has the greatest potential.
So today I'm excited to announce AIBOX.
AIBOX is a no-code AI app building platform paired with the App Store for AI that lets you monetize your AI tools.
The platform lets you build apps by linking together AI models like chatGPT, mid-journey and 11Labs, eventually will integrate with software like Gmail, Trello and Salesforce so you can use AI to automate every function in your organization.
To get notified when we launch and be one of the first to build on the platform, you can join the wait list at AIBOX.AI, the link is in the show notes.
We are currently raising a seed round of funding.
If you're an investor that is focused on disruptive tech, I'd love to tell you more about the platform.
You can reach out to me at jaden at AIBOX.AI, I'll leave that email in the show notes.
Welcome to the AI Chat podcast.
I'm your host, Jayden Schafer.
Today on the podcast, we have the pleasure of being joined by Rhea Mira, who is the VP of product at Aporia, a leading AI performance and observability platform.
Previously, he served as head of product at Arise AI, led some product initiatives centered around foundational language models and graph-based ML at Google AI, and played a pivotal role in the development of natural language models across a bunch of different industry applications at IBM Watson.
So Rhea is an alumnus of UC Berkeley electrical engineering and computer science class of 2014.
Welcome to the show today, Rhea.
Thanks so much for having me, Jayden.
How's it going?
I am doing fantastic.
You're calling in from Tel Aviv.
That's exciting.
A majority of our guests are from the United States.
So it's always nice to have people calling in from other places.
Tell us a little bit about your, I would just be curious, tell me a little bit about your background.
So you're in Tel Aviv now.
Are you originally from Tel Aviv?
What got you interested in AI and kind of this whole space?

Yeah.
So I was actually born and raised in Los Angeles, California.
So the Tel Aviv move was actually pretty spontaneous and as recent as a year ago, but yeah, born and raised in LA in the San Fernando Valley, went to Berkeley for university and then right after Berkeley, I actually moved to New York City and kind of all throughout, I started as a software engineer.
My first kind of stint was at NASA's Jet Propulsion Laboratory, working on the Mars Curiosity Rover and then kind of quickly after, through a number of academic courses, which I took, building AI adversarial Pac-Man agents and such, I became interested in making computers think, act and behave like humans and that kind of led me into machine learning and AI and it escalated fairly quickly.
Very cool.
So you said your first stint, you're working at NASA, super interesting.
Then you go over to Google.
How long were you at Google and I guess what were some of the things that you're working on over there?
Yeah, so I was actually at IBM's Watson Group before Google, but I was at Google.
I was in Google's research and machine intelligence division for about three years.
I worked on a team that was focused on a number of different graph convolutional networks or graph-based ML algorithms.
And what that means is it's really the core of, it's a core technology that can be used to power different applications from things like YouTube's Watch Next recommendation algorithm, all the way to fighting abuse across Gmail or determining what's NSFW content across search, so on and so forth.
It's pretty widely applicable and actually some of these algorithms even made it as far as broader alphabet.
So within Waymo's autonomous driving division and Verily, which is kind of Alphabet's healthcare and sciences department for different drug synthesis and cancer diagnosis.
So it's really kind of like a foundational layer on top of which you can build different applications.
Very, very interesting.
That's super cool.
And then my other question would be, yeah, tell us a little bit about what you're doing over at IBM with Watson.
It's kind of interesting because right now we see IBM really with Watson and Google being two of these major powerhouses when it comes to AI and the products they're putting out.
And we've seen a little bit about IBM and Watson kind of in the news.
I hear a lot more about Google, I feel like, and everything.
They're integrating maybe that's just a more brand recognition, but tell everyone a little bit about IBM and Watson and yeah, what you're doing there.
Yeah, so most people know IBM for kind of being a pioneer in AI, building this chess spot that Gary Kasparov and then subsequently beat Ken Jennings in Jeopardy.
But what I was actually working on was natural language models.

So everyone's familiar with the large language model Hype in present day.
I was working on much smaller and more fine tuned models for specific use cases and one of the more lucrative industries for a company that's generally B2B like IBM was actually legal and compliance.
So interesting.
So what might generally take a lawyer or a paralegal or a compliance officer hours if not hours on end or perhaps even days depending on the contract.
The objective was could you use a language model to extract firm specific language.
In other words, the redlining that lawyers do and you get a machine learning model to do just that.
Extract things like parties, buyer, supplier, differentiate between topics.
So obligations, rights, exclusions, disclaimers, all these different very niche entities and both syntactically and semantically.
And the reason that's important because we might touch on the topic of hallucinations but when it comes down to legal language, the difference between a multi-million dollar lawsuit and a bulletproof contract can be as minute as an Oxford comma.
So these models had to be extremely precise.
Okay.
Yeah, that sounds like the precision is really important and I'm sure this brings you a little bit to what you're doing today.
One other question that I'm curious about this is what kind of made you ship from IBM to Google?
What was that transition like?
Were you just seeing new opportunities or what was that move for?
Yeah, I guess for me it was a similar tenure at IBM's Watson Group.
I was there for approximately three years.
I actually kind of formally made the switch from software engineering to product management and really sat at that intersection of the business and the technology and I had a great time at IBM.
Nonetheless, it's a hundred year old company and Google seemed a bit more lean and agile.
Okay.
So I feel like I'm seeing this really interesting trend from you, right?
You're starting at NASA, so government, then you go to IBM, a hundred year old company, then you're going to Google a little bit more lean and then you decide to make the move to Aporia, which yeah, what was your thinking there?
This is obviously a new startup.
It's a lot less secure than a big, huge company like Google.
What was your thinking in that?
Yeah, great question.
Going from big corporations where oftentimes you can get a sense of a feeling of lost or being not driving enough impact as quickly as you'd like to.
I have a very extreme personality.
If I kind of go for something, it's 110% and so I felt like I built a breadth and depth

across my technical skill set as well as my product expertise and I made plenty of errors along the way and the beauty of working at a large corporation is that they generally have the resources for you to make those mistakes.
You screw up an experiment or perhaps take a little longer to push something to production.
Google's not going to go on, however, once you build that confidence, I think jumping into the world of startups and particularly in Tel Aviv and Israel, which is a startup nation, you kind of really put your skill set and your experience to the test.
Very frugal in terms of resources, very agile in terms of the timeline and features that you need to deliver and customers you've made promises to and so on and so forth.
Yeah, I totally see that and personally, I'm working on startups.
I love startups.
It's what I'm all about too, so definitely an exciting but different.
I previously also worked at a 100-year-old company, so definitely very different.
I want to get into what you guys are building at Aporia, but I had one other question before that which is, you mentioned born and raised in California, moved to Tel Aviv.
Was that for Aporia or did you move there first and find them after?
What did that look like for you?
Yeah, great question.
My parents were actually born and raised in Israel, not in the center of Tel Aviv but a little north.
For me, I kind of grew up visiting the country and was always kind of in awe of the number of startups and just the innovation that came out of a country the size of New Jersey and out of it.
Whenever I would come and visit, I was kind of immersed and baffled at the same time.
While I was at Google, I would make several trips here for work.
The Tel Aviv office is one of the leaders in terms of AI and innovation.
For me, it was a combination of both personal exploration as well as professional development.
You moved and then you found Aporia after that while you were living there?
Yes.
Okay.
Very cool.
Super cool.
Honestly, I think that's a really interesting story for a lot of people listening.
The place you want to go and you'll find an amazing company to work.
You don't just have to live exactly where you may perhaps have a job today or whatnot.
Very cool story.
Tell us a little bit about Aporia and the problems you guys are currently solving for customers.
I would be happy to.
I guess one thing that really led me here was the fact that while being at these major corporations with kind of leading AI and ML organizations, I kind of saw the same challenges over and over again, largely unsolved, despite the resources, despite the team size and even the success of the models.

The traditional machine learning lifecycle starts at first off identifying the objective of what you want your model to do and then going from data collection and cleansing to feature engineering to training your model, running different experiments.
But once you launch that model into production, in other words, the real world, it is very difficult to gauge its performance and furthermore, to know when something goes wrong, how to fix it.
And that's exactly what Aporia is doing right now.
We are kind of a use case agnostic AI performance and observability platform.
Now that's a mouthful.
So I'll kind of break it down into maybe a few pillars.
You're an organization, you've got machine learning models.
These models are not for the sake of saying your company does AI.
They actually affect core business metrics or KPIs.
So a common example is you're at your favorite tech store, brick and mortar, or maybe you're even ordering from Amazon, but you've been saving up and you transact on something that's maybe higher than your usual spend and you get that text message we are all familiar with.
Is this really you or is this wrong?
That's a machine learning model.
And so the number of what Aporia does is we actually allow you to upload different models into your own data lake or warehouse and we connect directly to that.
And we empower kind of, I want to say three and soon to be four pillars.
The first is centralized visibility.
So a single kind of bird's eye or centralized view of your models, their activity, performance, their overall health in terms of data quality, drifts and any other issues that you might encounter.
The reason I mentioned that the example fraud use case is because for a big bank or a credit card company, those types of models make very powerful business decisions, whether to allow a transaction and if it's a false negative, in other words, if it's a fraudulent transaction that goes under the radar, that's a chargeback or a loss for the company.
So what Aporia does or its second pillar is proactive monitoring.
You want to be alerted when something goes wrong with your model before that downstream business KPI is affected.
So if for example, you're seeing a spike in false negative fraud transactions, people are calling in and saying, hey, chargeback my credit card, that was me.
You want to be proactively alerted before that drift or that we call drift kind of like a slow bleeding failure, change over time, before it really affects your business.
That's the second pillar, which is monitoring.
The third pillar that Aporia offers is root cause analysis or in other words, the why.
So monitoring kind of gives you what happened, hey, you're experiencing this performance degradation, you're experiencing this drift, there might be a data quality issue, but why or the root cause analysis allows you to quickly discover what is the reason that it happened and gives you insights or actionable recommendations on how to fix those issues in production.

I'll kind of pause there because that was a bit of a rant, but we're actually working on some fairly new features for the large language model space and the hype that the world is seeing as we speak today.
Yeah, well, to be honest, that sounds like some really fascinating stuff you're working on and I see some serious value, especially for the enterprise, but what you're kind of alluding to there is what I'd be interested in asking is kind of like, so you guys have these pillars that you have in Aporia that it currently does, what are some of the exciting things you guys are planning on or you see a need for in the future so you're building today?
Yeah, I think the natural course of the conversation leads us to the widespread applications built on top of large language models and with the movement of open source models and open source data sets and kind of pre-trained off-the-shelf models that you can use, you no longer need to be a data scientist to leverage ML models, right?
Any software engineer or even homegrown coder can go ahead and pick up and off the shelf model and bang away at the keyboard and call the API and get some pretty fascinating results, whether it's sentiment analysis or knowledge answering or question answering.
Now the challenges come when you're again building kind of a real world application, you trust this large language model built on 500 billion parameters, whatever the GPT-7 will be, right?
Like, you've got models that are built on hundreds of billions of parameters in the future, we might see trillions of parameters, right?
And you know, insimously, it's very difficult to understand why your model acted in the way that it did.
What led to this prediction, right?
Whether it's a hallucination or whether it's a concept and for that reason, this concept of AI guardrails, which we're working on and really building out in a sophisticated way is something I'm really excited about, being able to detect and prevent AI hallucinations in a continuous and iterative fashion so that, you know, these types of blunders, right?
And there are real-world examples, this isn't kind of like a made-up challenge, right?
I think the release of Google's BARD, what happened was, you know, Sundar hopped on stage and actually, you know, flaunted BARD during this launch and it made a blunder.
I mean, I think it was the Hubble telescope, it kind of like launched that and Google's stock flooded it, right?
And so, yeah, 100 billion dollars.
100 billion dollars lost in a snap.
So how do you detect these types of hallucinations?
How do you mitigate these challenges?
And furthermore, you know, how do you prevent them going forward?
How do you continuously monitor and develop a feedback loop that's crucial in identifying these types of issues?
Yeah.
So, I mean, that's my question, like, how do you do that?
How do you, you know, prevent these hallucinations?

Obviously, this is a big issue, 100 billions of dollars on the line.
I mean, a side note, it's kind of, I feel like that's Google's fault because they had like a, they should have had like, it was literally their marketing team should have just fact-checked the slides they had up, but, you know, that's another problem.
But yeah, like, how do you, how do you fix this?
This is a big, a big problem, especially for, you know, the end consumer that uses chat GPT for small projects, not a big deal.
They'll fact-check it, they'll figure it out.
But the real big problem is corporations and the enterprise, like, they can't have these mistakes, right?
This is incredibly important to them.
So, so yeah, that's the question, like, how do you, how do you stop that?
Yeah, I think, you know, it's a very loaded question.
And I'll do my best to kind of gloss over some of the methods that, that we're using and building out in, in a way that I can elaborate on.
Yeah.
You know, that there are generally kind of three ways of evaluating a large language model application or an application built on any sort of language model.
The first are a series of task-based metrics.
So, you know, for those who are interested, there's a Stanford paper called Helm, which stands for Holistic Evaluation of Language Models.
And there you can kind of see, I don't know, like, 50 plus difference LLM models evaluated across a number of different scenarios, as well as, you know, different metrics, whether it's exact match or quasi exact match, et cetera.
So depending on the task, there are metrics that we as, you know, a species of data scientists and statisticians have developed to measure accuracy, precision recall, et cetera.
That's kind of the first method.
Now, the second method actually falls on user feedback.
So you may have seen or not seen, you know, some degree of, you know, in chat GPT on Open AI's site, right?
You've got the thumbs up or thumbs down or, you know, rate response or which response is better.
So what you're actually doing is you are a human reinforcement kind of feedback that's helping fine tune that model.
And so being able to leverage a combination of task based metrics, as well as user feedback, generally segmented by, you know, different features or some sort of segmentation, you can go ahead and make educated guesses as to the probability of a response being relevant, being truthful, being harmful, being a hallucination, and so on and so forth.
And actually maybe the third and perhaps most interesting but controversial method is actually having an LLM evaluate an LLM, which is interesting, right?
Because on the one hand, you've got maybe, you know, again, I'll just kind of throw 500 billion parameter model that's spewing questions and responses to everything from, you know, summarize this article to, you know, narrate this show in Shakespearean language, right?

And read, you know, your job is to really detect those types of topics.
And again, it's bias or toxicity or hallucinations, etc.
And there's kind of this novel movement of using smaller, more fine tuned models to evaluate their responses.
And they're actually language models themselves, which is, you know, trying to like, you know, solve a problem with the root of what's causing the problem in and of itself.
Yeah.
But yeah, I'll kind of pause there and get your thoughts on it.
Yeah, well, I was just going to say like, so obviously, you mentioned it's kind of controversial using an AI model to evaluate an AI model.
Like why do you, what do you, I guess, what's the core of the controversy there?
Like why do you think it is the most controversial?
Why do you think people have like an aversion to that?
Like, because theoretically it's like, oh, it's one tool making another tool better.
This tool was specifically designed to do this.
What do you think people are really concerned about here?
You know, I think the concerns around AI and their, you know, essentially the transparency of their predictions ultimately falls under the training data itself, right?
Because your models will only ever be as good and as diverse as your training data, which you hope is representative of real world scenarios that your model will encounter.
And so, you know, there's this concept of not just measuring the output of a model through its predictions and through, you know, some degree of a performance metric measurement, meaning, you know, maybe you have a sample of your data or your model's predictions actually being vetted by humans and evaluated by humans.
But there is also a degree of, I guess what we'd call like, auditing that's very important along the way.
Yeah.
So like, it actually starts from the data you've collected.
Who and how was it labeled?
Is it representative?
Is it diverse?
Is it, does it implicitly mitigate against bias?
You know, it is the feedback that you're collecting even bias, right?
Are all the users that you're using to evaluate from the same group or segment or even geographical location?
Yeah, yeah, yeah.
So, you know, it falls under these kind of responsible AI principles, which I think are, you know, it's a huge rabbit hole and it's extremely important, but it's something that, you know, big and small and every type of organization or every size of, in between, you know, organizations are trying to solve this.
And at the end of the day, the end users, they uphold us, whether it's Aporia or Google Bard or ChatGPT and OpenAI uphold, you know, these models to a high standard of excellence.
Right?

And I think, you know, it's one of the more exciting areas of research and development today.
Totally.
Very, very interesting.
Well, Rea, this has been absolutely fascinating to kind of hear your opinions and pick your brain and see what you guys are building right there.
Obviously there's some big, like, issues and things we're still grappling with in AI, but it's been amazing to have you on the show.
I'm wondering, you know, before we wrap this up, I'm wondering if you have a piece of advice that you could give to perhaps software developers or other people implementing AI into their companies right now.
What's a piece of advice you feel like you could give them?
Wow.
I've got several pieces.
If I had to pick just one, you know, if you are working on building an AI application, you know, whether you're using an off-the-shelf model or fine-tuning your own, I really do want to encourage you to either, you know, kind of build in-house or use existing off-the-shelf tools.
You know, you don't have to reinvent the wheel these days.
Your models, performance, and avoid creating or reinforcing bias, you know, unfair bias.
I think it's, you know, we as engineers and product folks and entrepreneurs have to be accountable for the technology that we put out there in the world.
And for that reason, you know, I do want to encourage some degree of responsibility when it comes to launching these AI applications.
So, you know, for that, I'd be ever grateful if you took away from this conversation.
That's a great piece of advice.
Really appreciate that.
If people want to, you know, contact you or they want to learn more about Aporia, maybe try it out.
What's the best way for them to do that?
Yeah, you can reach out to me at rea, my first name, R-E-A-H dot AI, and you'll find all my info there.
Okay.
Amazing.
And I will leave a link to Aporia in the comments or in the description as well.
So people will be able to find that.
But R-E-A, thank you so much for coming on the show today.
To the listener, thank you so much for listening to the AI Chat podcast.
Make sure to rate us wherever you listen to your podcasts and have an amazing rest of your day.
If you are looking for an innovative and creative community of people using ChatGPT, you need to join our ChatGPT creators community.
I'll drop a link in the description to this podcast.

We'd love to see you there where we share tips and tricks of what is working in ChatGPT.
It's a lot easier than a podcast as you can see screenshots, you can share and comment on things that are currently working.
So if this sounds interesting to you, check out the link in the comment.
We'd love to have you in the community.
Thanks for joining me on the OpenAI podcast.
It would mean the world to me if you would rate this podcast wherever you listen to your podcasts and I'll see you tomorrow.