KARINA NGUYEN

LENNY'S PODCAST

BILINGUAL TRANSCRIPT

ORIGINAL BY

Lenny Rachitsky

@lennysan • x.com/lennysan

ANALYSIS BY

@Penny777 • x.com/penny777

Generated: 2026-01-16

Karina Nguyen - 双语对照

Lenny's Podcast: Karina Nguyen (OpenAI Researcher) - Bilingual Transcript

(00:00:00) Lenny Rachitsky

English:

Not only are you working at the cutting edge of AI and LLMs, you're actually building the cutting edge.

中文翻译:

你不仅是在人工智能（AI）和大语言模型（LLM）的最前沿工作，你实际上正在亲手打造这个最前沿。

(00:00:06) Karina Nguyen

English:

When I first came to Anthropic and I was like, "Oh my God, I really love front-end engineering." And then the reason why I switched to research is because I realized, "Oh my God, Claude is getting better at front-end. Claude is getting better at coding. I think Claude can develop new apps."

中文翻译:

当我刚加入 Anthropic 时，我觉得：“天哪，我真的很喜欢前端工程。”后来我转向研究领域的原因是，我意识到：“天哪，Claude 的前端能力越来越强了。Claude 的编程能力越来越强了。我觉得 Claude 甚至可以开发新的应用程序了。”

(00:00:20) Lenny Rachitsky

English:

What skills do you think will be most valuable going forward for product teams, in particular?

中文翻译:

你认为在未来，特别是对于产品团队来说，哪些技能将是最有价值的？

(00:00:26) Karina Nguyen

English:

Creative thinking and you kind of want to generate a bunch of ideas and filter through them and not just build the best product experience. I think it's actually really, really hard to teach the model how to be

aesthetic or really good visual design or how to be extremely creative in the way they write.

中文翻译:

创造性思维。你需要产生大量的想法并从中筛选，而不仅仅是构建最佳的产品体验。我认为，要教会模型如何具备审美观、如何做极其出色的视觉设计，或者如何在写作方式上极具创意，实际上是非常非常困难的。

(00:00:42) Lenny Rachitsky

English:

What do you think people most misunderstand about how models are created?

中文翻译:

你认为人们对于模型是如何创建的，最大的误解是什么？

(00:00:46) Karina Nguyen

English:

When you taught the model, some of the self-knowledge of you actually don't have a physical body to operate in the physical world, the model would get extremely confused.

中文翻译:

当你教给模型一些自我认知，比如“你实际上没有在物理世界中操作的实体身体”时，模型会变得极其困惑。

(00:00:58) Lenny Rachitsky

English:

Today my guest is Karina Nguyen. Karina is an AI researcher at OpenAI where she helped build Canvas, tasks, the o1 chain-of-thought model and more. Prior to OpenAI, she was at Anthropic where she led work on post-training and evaluation for the Claude 3 models, built a document upload feature with 100K context windows and so much more. She was also an engineer at New York Times, was a designer at Dropbox and at Square. It's very rare to get a glimpse into how someone working on the bleeding edge of AI and LLMs operates and how they think about where things are heading.

中文翻译:

今天的嘉宾是 Karina Nguyen。Karina 是 OpenAI 的一名 AI 研究员，她参与构建了 Canvas、任务功能（tasks）、o1 思维链（chain-of-thought）模型等。在加入 OpenAI 之前，她在 Anthropic 工作，负责 Claude 3 模型的后训练（post-training）和评估工作，构建了具有 10 万上下文窗口的文档上传功能等等。她还曾是《纽约时报》的工程师，以及 Dropbox 和 Square 的设计师。能够一窥在 AI 和大语言模型最前沿工作的人是如何运作的，以及他们如何看待未来的走向，是非常难得的机会。

(00:01:31) Lenny Rachitsky

English:

In our conversation, we talk about how teams that OpenAI operate and build products, what skills she thinks you should be building as AI gets smarter, how models are created, why synthetic data will allow models to keep getting smarter and why she moved from engineering to research after realizing how

good LLMs are going to be at coding. If you enjoy this podcast, don't forget to subscribe and follow it in your favorite podcasting app or YouTube. It's the best way to avoid missing feature episodes and it helps the podcast tremendously. With that, I bring you Karina Nguyen.

中文翻译:

在我们的对话中，我们讨论了 OpenAI 的团队如何运作和构建产品，她认为随着 AI 变得越来越聪明，你应该培养哪些技能，模型是如何创建的，为什么合成数据（synthetic data）能让模型持续进化，以及为什么她在意识到大语言模型的编程能力将变得如此强大后，从工程转向了研究。如果你喜欢这个播客，别忘了在你的播客应用或 YouTube 上订阅和关注。这是避免错过未来节目的最好方式，也对播客有很大帮助。下面，让我们欢迎 Karina Nguyen。

(00:02:02) Lenny Rachitsky

English:

This episode is brought to you by Enterpret. Enterpret unifies all your customer interactions from Gong calls to Zendesk tickets to Twitter threads to app store reviews, and makes it available for analysis. It's trusted by leading product orgs like Canva, Notion, Loom, Linear, monday.com, and Strava, to bring the voice of the customer into the product development process, helping you build best-in-class products faster. What makes Enterpret special is its ability to build and update customer-specific AI models that provide the most granular and accurate insights into your business, connect customer insights to revenue and operational data in your CRM or data warehouse to map the business impact of each customer need and prioritize confidently, and empower your entire team to easily take action on use cases like win-loss analysis, critical bug detection and identifying drivers of churn with Enterpret's AI system, Wisdom.

中文翻译:

本集节目由 Enterpret 赞助。Enterpret 统一了你所有的客户互动——从 Gong 通话到 Zendesk 工单，从 Twitter 帖子到应用商店评论，并使其可用于分析。它深受 Canva、Notion、Loom、Linear、monday.com 和 Strava 等领先产品组织的信任，将客户的声音带入产品开发过程，帮助你更快地构建一流产品。Enterpret 的特别之处在于它能够构建和更新针对特定客户的 AI 模型，提供关于你业务的最细致、最准确的洞察；将客户洞察与 CRM 或数据仓库中的收入和运营数据连接起来，以映射每个客户需求的业务影响并自信地确定优先级；并通过 Enterpret 的 AI 系统 Wisdom，赋能你的整个团队轻松应对盈亏分析、关键漏洞检测和识别流失驱动因素等用例。

(00:02:53) Lenny Rachitsky

English:

Looking to automate your feedback loops and prioritize your roadmap with confidence, like Notion, Canva and Linear? Visit E-N-T-E-R-P-R-E-T.com/Lenny to connect with the team and to get two free months when you sign up for an annual plan. This is a limited time offer. That's Enterpret.com/Lenny. This episode is brought to you by Vanta. And I am very excited to have Christina Cacioppo, CEO and co-founder Vanta, joining me for this very short conversation.

中文翻译:

想要像 Notion、Canva 和 Linear 一样自动化你的反馈循环并自信地规划路线图吗？访问 Enterpret.com/Lenny 联系团队，并在注册年度计划时获得两个月的免费试用。这是一个限时优惠。网址是 Enterpret.com/Lenny。本集节目由 Vanta 赞助。我非常高兴邀请到 Vanta 的首席执行官兼联合创始人 Christina Cacioppo 加入这段简短的对话。

(00:03:22) Christina Cacioppo

English:

Great to be here. Big fan of the podcast and the newsletter.

中文翻译:

很高兴来到这里。我是这个播客和新闻通讯（newsletter）的忠实粉丝。

(00:03:25) Lenny Rachitsky

English:

Vanta is a longtime sponsor of the show, but for some of our newer listeners, what does Vanta do and who is it for?

中文翻译:

Vanta 是本节目的长期赞助商，但对于一些新听众来说，Vanta 是做什么的，它是为谁服务的？

(00:03:32) Christina Cacioppo

English:

Sure. So we started Vanta in 2018. Focused on founders, helping them start to build out their security programs and get credit for all of that hard security work with compliance certifications like SOC 2 or ISO 27001 today, we currently help over 9,000 companies, including some startup household names like Atlassian, Ramp and LangChain start and scale their security programs and ultimately build trust by automating compliance, centralizing GRC, and accelerating security reviews.

中文翻译:

好的。我们在 2018 年创立了 Vanta。我们专注于创始人，帮助他们开始建立安全计划，并通过 SOC 2 或 ISO 27001 等合规认证，让他们辛苦的安全工作获得认可。目前，我们帮助超过 9,000 家公司，包括 Atlassian、Ramp 和 LangChain 等知名初创公司，启动并扩展其安全计划，最终通过自动化合规、集中化 GRC（治理、风险与合规）和加速安全审查来建立信任。

(00:04:04) Lenny Rachitsky

English:

That is awesome. I know from experience that these things take a lot of time and a lot of resources and nobody wants to spend time doing this.

中文翻译:

太棒了。我从经验中知道，这些事情需要花费大量的时间和资源，而且没人想把时间花在这些事情上。

(00:04:10) Christina Cacioppo

English:

That is very much our experience, but before the company and to some extent during it. But the idea is with automation, with AI, with software, we are helping customers build trust with prospects and

customers in an efficient way. And our joke, we started this compliance company so you don't have to.

中文翻译:

这正是我们的体会，在公司成立之前以及在某种程度上成立期间都是如此。但我们的理念是利用自动化、AI 和软件，帮助客户以高效的方式与潜在客户和现有客户建立信任。我们常开玩笑说：我们创办这家合规公司，就是为了让你们不必再为合规操心。

(00:04:26) Lenny Rachitsky

English:

We appreciate you for doing that. And you have a special discount for listeners, they can get $1,000 off Vanta at Vanta.com/Lenny, that's V-A-N-T-A.com/Lenny for $1,000 off Vanta. Thanks for that, Christina.

中文翻译:

感谢你们所做的一切。你们还为听众提供了特别折扣，在 Vanta.com/Lenny 可以获得 1,000 美元的优惠。网址是 V-A-N-T-A.com/Lenny。谢谢你，Christina。

(00:04:41) Christina Cacioppo

English:

Thank you.

中文翻译:

谢谢。

(00:04:45) Lenny Rachitsky

English:

Karina, thank you so much for being here. Welcome to the podcast.

中文翻译:

Karina，非常感谢你能来。欢迎来到播客。

(00:04:48) Karina Nguyen

English:

Thank you so much, Lenny, for inviting me.

中文翻译:

非常感谢 Lenny 邀请我。

(00:04:50) Lenny Rachitsky

English:

I'm very excited to have you here because not only are you working at the cutting edge of AI and LLMs, you're actually building the cutting edge of AI and LLMs. You recently launched this feature, which

basically... the first agent feature of OpenAI. I also just did this survey, I don't know if you know about this. I did a survey of my readers and asked them what tools do you use every day in your work and most use? And ChatGPT was number one, above Gmail, above Slack, above anything else. 90% of people said they use ChatGPT regularly.

中文翻译:

我非常激动能邀请到你，因为你不仅是在 AI 和大语言模型的最前沿工作，你实际上正在亲手打造它。你最近发布了这个功能，基本上是……OpenAI 的第一个智能体（agent）功能。我最近还做了一个调查，不知道你是否听说过。我调查了我的读者，问他们每天工作中必用且最常用的工具是什么？ChatGPT 排名第一，超过了 Gmail、Slack 和其他任何工具。90% 的人表示他们经常使用 ChatGPT。

(00:05:23) Karina Nguyen

English:

That's quite good.

中文翻译:

那相当不错。

(00:05:23) Lenny Rachitsky

English:

It's absurd. It wasn't around two years ago.

中文翻译:

这简直不可思议。两年前它甚至还不存在。

(00:05:25) Karina Nguyen

English:

Yeah.

中文翻译:

是的。

(00:05:26) Lenny Rachitsky

English:

Also, we're recording this the week that OpenAI announced Stargate, which is this half trillion dollar investment in AI infrastructure. So there's just a lot happening constantly in AI and you have a really unique glimpse into how things are working, where things are going, how work gets done. So I have a lot of questions for you. I want to talk about how you operate and how you work at OpenAI, where you think things are going, what skills are going to matter more and less in the future, and also just where things are going broadly. So how does that sound?

中文翻译:

此外，我们录制这期节目的时候，正好是 OpenAI 宣布 Stargate（星际门）计划的那一周，这是一项耗资五千亿美元的 AI 基础设施投资。AI 领域不断有大事发生，而你对事情如何运作、未来走向以及工作如何开展有着非常独特的视角。所以我有很多问题想问你。我想聊聊你在 OpenAI 是如何运作和工作的，你认为未来的趋势是什么，哪些技能在未来会变得更重要或不那么重要，以及大方向上的发展。你觉得怎么样？

(00:05:55) Karina Nguyen

English:

Sounds great. Thank you so much. Yeah, I was extremely lucky to join early days Anthropic and learned a lot of things there. And I joined OpenAI around eight months ago. So, yeah, I'm excited to dive more in into-

中文翻译:

听起来很棒。非常感谢。是的，我非常幸运能在早期加入 Anthropic 并学到了很多东西。大约八个月前我加入了 OpenAI。所以，我很兴奋能深入探讨——

(00:06:11) Lenny Rachitsky

English:

Okay, I'm going to definitely ask you about the differences between those, but I want to start more technical and just dive right in. I want to talk about model training. People always hear about models being trained, these big models, how much data takes, how long it takes, how much money toss it takes, how we're running out of data, which I want to talk about. Let me just ask you this question. What do you think people most misunderstand about how models are created?

中文翻译:

好的，我肯定会问你这两家公司的区别，但我先从技术层面开始，直接切入正题。我想聊聊模型训练。人们总听说模型在训练，这些大模型需要多少数据，花多长时间，耗费多少资金，以及我们正面临数据枯竭的问题——这也是我想聊的。我先问你这个问题：你认为人们对于模型是如何创建的，最大的误解是什么？

(00:06:36) Karina Nguyen

English:

Model training is more an art than a science. And in a lot of ways we, as model trainers, think a lot about data quality. It's one of the most important things in model training is like how do you ensure the highest quality data for certain interaction model behavior that you want to create? But the way you debug models is actually very similar the way you debug software. So one of the things that I've learned early days at Anthropic was we've discovered especially this Claude 3 training, when you taught the model some of the soft-knowledge of, "Hey, you actually don't have a physical body to operate in the physical world." But then at the same time you had data that taught the model some of the function calls, which is like, "This is how you set the alarm."

中文翻译:

模型训练更多是一门艺术，而非科学。在很多方面，作为模型训练者，我们非常关注数据质量。模型训练中最重要的事情之一就是：如何确保最高质量的数据，以创造出你想要的特定交互模型行为？但调试模型的方式实际上与调试软件非常相似。我在 Anthropic 早期学到的一件事是，特别是在 Claude 3 的训练中，我们发现当你

教给模型一些软知识，比如“嘿，你实际上没有物理身体在现实世界中操作”，但与此同时，你又有数据教给模型一些函数调用（function calls），比如“这是你设置闹钟的方法”。

(00:07:30) Karina Nguyen

English:

And so the model would get extremely confused about whether it can set an alarm, but it doesn't have a body in the physical world. So it's like the model gets confused and sometimes it'll over accuse. So sometimes it says, "Look, I don't know. Sorry, I cannot help you." And so there is always a balance trade off between how do you make the model to be more helpful for users, but also not being harmful in other scenarios. And so it's always about how do you make the model more robust and operate across a variety of diverse scenarios.

中文翻译:

于是模型就会变得极其困惑：它到底能不能设置闹钟？但它在物理世界中又没有身体。模型会感到混乱，有时甚至会“过度拒绝”。它有时会说：“听着，我不知道。抱歉，我帮不了你。”所以，在如何让模型对用户更有帮助，同时又不在其他场景中产生危害之间，总是存在一种权衡。这始终关乎如何让模型更具鲁棒性（robust），并在各种不同的场景中正常运行。

(00:08:09) Lenny Rachitsky

English:

That is so funny. I never thought about that. Most of the data that it's trained on is kind of assuming it's like a human describing the world and how they operate. It assumes there's a body and you could do things, and the model is told you don't have a body.

中文翻译:

这太有趣了。我从未想过这一点。它训练所用的大部分数据都假设它像人类一样描述世界及其运作方式。数据假设有一个身体可以做事，但模型却被告知你没有身体。

(00:08:20) Karina Nguyen

English:

Yeah.

中文翻译:

是的。

(00:08:21) Lenny Rachitsky

English:

Okay. I want to talk a little bit about data while we're on this topic. I know you have strong opinions here. There's this meme that models are going to stop getting smarter because they're running out of data. They're trained in a large part on the internet and there's only one internet and they've already been trained on it, what more can you show them about the world? And there's this trend of synthetic data,

this term synthetic data. What is synthetic data? Why do you think it's important? Do you think it's going to work?

中文翻译:

好。既然聊到这个话题，我想谈谈数据。我知道你在这方面有很强的见解。现在有一种说法（meme），认为模型会停止变得更聪明，因为数据快用完了。它们很大程度上是在互联网上训练的，而互联网只有一个，它们已经训练过了，你还能给它们展示什么关于世界的新东西呢？于是出现了“合成数据”（synthetic data）这个趋势。什么是合成数据？为什么你认为它很重要？你认为它会奏效吗？

(00:08:47) Karina Nguyen

English:

I think there are two questions here. We can unpack one at a time. But people say we are hitting the data wall. I think people think more in the terms of pre-trained large models that are trained on the entire internet to predict the next token. But what actually the model is learning during that process is actually how do you compress the compression algorithm here? The model learns to compress a lot of knowledge and it learns how to model the world. So the next prediction of the word, like, "Teach me how to drive," basically. And you only have a few words that will match that, a car. So the model actually learns about the world in itself. So it's like it's modeling human behavior, sometimes it's modeling... And when you talk to pre-trained models which are very, very large, they're actually extremely diverse and extremely creative because you can talk to almost any Reddit user through a pre-trained model.

中文翻译:

我认为这里有两个问题，我们可以逐一拆解。人们说我们正面临“数据墙”。我认为人们更多是从预训练大模型的角度来考虑的，这些模型在整个互联网上训练以预测下一个标记（token）。但模型在这个过程中实际学习的是如何压缩——这本质上是一个压缩算法。模型学会了压缩大量知识，并学会了如何对世界建模。比如预测下一个词，“教我如何驾驶”，基本上只有几个词能匹配，比如“汽车”。所以模型实际上学习的是世界本身。它在模拟人类行为，有时在模拟……当你与非常大的预训练模型交谈时，它们实际上极其多样化且极具创意，因为通过预训练模型，你几乎可以与任何 Reddit 用户交谈。

(00:09:56) Karina Nguyen

English:

But I think what's happening right now with new paradigm of o1 series is that the scaling in post-training itself is not hitting the wall. And that's because basically we went from raw data sets from pre-trained models to infinite amount of tasks that you can teach the model in the post-training world via reinforcement learning. So any task, for example, how to search the web, how to use the computer, how to write, wow, all sorts of tasks that you trying to teach the model all the different skills. And that's why we're saying there's no data wall or whatever, because there will be infinite amount of tasks and that's how the model becomes extremely super intelligent. And we are actually getting saturated in all benchmarks.

中文翻译:

但我认为，随着 o1 系列这种新范式的出现，后训练（post-training）阶段的扩展并没有遇到瓶颈。这是因为我们基本上从预训练模型的原始数据集，转向了在后训练阶段通过强化学习（reinforcement learning）教给模型的无限任务。例如，任何任务：如何搜索网页、如何使用电脑、如何写作，哇，各种各样的任务，你试图教给模型所有不同的技能。这就是为什么我们说没有所谓的“数据墙”，因为任务是无限的，这就是模型变得极其超智能的方式。实际上，我们在所有基准测试（benchmarks）中都快达到饱和了。

(00:10:52) Karina Nguyen

English:

So I think the bottleneck is actually in evaluations that we don't have all the frontier, like evals like, I don't know, GPQA, which is a Google-proof question answering, PhD level intelligence. The benchmark is getting to, I don't know, more than 60, 70%, which is what PhD gets. So it's literally hitting the wall in like evals.

中文翻译:

所以我认为瓶颈实际上在于评估（evaluations）。我们没有足够的尖端评估手段，比如 GPQA（谷歌搜不到答案的问答测试），它代表博士级别的智能。基准测试的分数已经达到了 60%、70% 以上，这正是博士能达到的水平。所以，真正“撞墙”的是评估手段。

(00:11:19) Lenny Rachitsky

English:

I want to follow both those threads. So the first is on this idea of synthetic data. Is a simple way to understand it, that the models are generating the data that future models are trained on and you ask it to generate all these ways of doing stuff, all these tasks as you described, and then the newer models trained on this data that the previous model generated?

中文翻译:

我想顺着这两个思路聊聊。首先是关于合成数据的想法。一种简单的理解方式是不是：模型生成数据，然后未来的模型基于这些数据进行训练？你让它生成各种做事的方法、各种你描述的任务，然后新模型就在旧模型生成的这些数据上进行训练？

(00:11:39) Karina Nguyen

English:

Some tasks are synthetically curated. So this is an active research area is how can you synthetically construct new tasks with models to learn. Sometimes when you develop products, you get a lot of data from the product and user feedback and you can use that data too in this cross-training world. Sometimes you still want to use human data because actually some of the tasks can be really, really hard to teach. Experts only know certain knowledge about some chemicals or biological knowledge, so you actually need to tap into the experts' knowledge a lot. So yeah, I think to me synthetic data training is more for product... It's a rapid model iteration for similar product outcomes. And we can dive more into it, but the way we made Canvas and tasks and new product features for ChatGPT was mostly done by synthetic training.

中文翻译:

有些任务是人工合成策划的。这是一个活跃的研究领域：如何利用模型合成地构建新任务供其学习。有时在开发产品时，你会从产品和用户反馈中获得大量数据，你也可以在交叉训练中使用这些数据。有时你仍然需要使用人类数据，因为有些任务确实非常难以教授。只有专家才了解某些化学或生物知识，所以你实际上需要大量挖掘专家的知识。所以，对我来说，合成数据训练更多是为了产品……它是为了实现类似产品结果的快速模型迭代。我们可以深入探讨，但我们为 ChatGPT 制作 Canvas、任务功能和其他新产品功能的方式，主要就是通过合成训练完成的。

(00:12:52) Lenny Rachitsky

English:

Let's actually get into that. That's really interesting. I want to talk about evals, but let's follow that thread. So talk about how this helped you create Canvas.

中文翻译:

那我们深入聊聊这个。这非常有意思。我想聊聊评估（evals），但先顺着这个思路：谈谈这如何帮助你创建了 Canvas。

(00:12:56) Karina Nguyen

English:

So when I first came to OpenAI, I really had this idea of, "Okay, it would be really cool for ChatGPT to actually change the visual interface but also change the way it is with people." So going from being a chatbot to more of a collaborative agent, and the collaborator is a step towards more genetic systems that become innovators ultimately. And the entire team of applied engineers, designers, products, research got formed in the air almost out of nothing. It's just like a collection of people who just got together and we rapidly started iterating with each other.

中文翻译:

当我刚来到 OpenAI 时，我真的有这样一个想法：“如果 ChatGPT 能改变视觉界面，同时也改变它与人相处的方式，那就太酷了。”也就是从一个聊天机器人转变为一个更具协作性的智能体（collaborative agent），而这种协作是迈向最终成为创新者的更通用系统的一步。于是，由应用工程师、设计师、产品经理和研究人员组成的整个团队几乎凭空组建了起来。就像一群志同道合的人聚在一起，我们开始快速地互相迭代。

(00:13:46) Karina Nguyen

English:

Actually Canvas is one of the... I would say the first project at OpenAI, where researchers and applying engineers started working together from the very beginning of the product development cycle. And I think there's a lot of things that we have learned on the way, but I definitely came with the mindset of, "We need to do a really rapid model situation such that it would be much easier for engineers to work with the latest model possible, but also learn from user feedback or early internal dog food. How do we improve the model very rapidly?"

中文翻译:

实际上，Canvas 是 OpenAI 的……我会说是第一个项目，研究人员和应用工程师从产品开发周期的最开始就一起工作。我想我们在过程中学到了很多东西，但我当时确实带着这样一种心态：“我们需要进行非常快速的模型迭代，这样工程师就能更容易地使用最新的模型，同时也能从用户反馈或早期的内部试用（dogfooding）中学习。我们如何极速改进模型？”

(00:14:28) Karina Nguyen

English:

And it's really hard to kind of like figure out how people... when you deploy a product, how people would be able to use it. And so the way you synthetically train the model is physically figuring out what are the

most core behaviors that you wanted the product feature to do. And for Canvas, for example, it came down to three main behaviors. It was how do you trigger Canvas for prompts like, "Write me a long essay," when the user intention is mostly iterating over long documents? Or, "Write me a piece of code," or when to not trigger Canvas for prompts like, "Can you tell me more about President..." I don't know, some of the general questions. So you don't want to trigger Canvas because the user intention is mostly getting answer, not necessarily iterate over the long document.

中文翻译:

当你发布一个产品时，很难预料人们会如何使用它。所以，合成训练模型的方法就是从物理上搞清楚你希望产品功能实现的内核行为是什么。以 Canvas 为例，它归结为三个主要行为。第一，如何针对“给我写一篇长文章”这样的提示词触发 Canvas？因为此时用户的意图主要是迭代长文档。或者“给我写一段代码”。以及什么时候*不*触发 Canvas，比如“你能告诉我更多关于总统……”之类的一般性问题。在这种情况下，你不希望触发 Canvas，因为用户的意图主要是获得答案，而不是迭代长文档。

(00:15:28) Karina Nguyen

English:

The second behavior is how do we teach the model to update the document when the user asks? So one of the behaviors that we taught the model is actually have some agency and autonomy to literally go to the document and select specific sections and either delete it or edit, so highlight it and rewrite certain sections. Sometimes the user would just say, "Change the second paragraph to be something friendlier," and we would have to teach the model to literally find the second paragraph in the document and change it to a friendly tone. So basically you teach both how to trigger edit itself, but also how do you teach the model to get higher quality edit for the document?

中文翻译:

第二个行为是：当用户要求时，我们如何教会模型更新文档？我们教给模型的行为之一是让它具备一定的自主性，能够直接进入文档，选择特定部分，然后删除或编辑——也就是高亮并重写某些部分。有时用户会说：“把第二段改得更友好一点”，我们就必须教会模型准确找到文档中的第二段，并将其改为友好的语气。所以，你既要教它如何触发编辑本身，也要教它如何为文档提供更高质量的编辑。

(00:16:21) Karina Nguyen

English:

In case of coding, for example, there's also the question of how good the model is of completely rewriting the document, versus having a very specific target edits. So that's another layer of decision boundary within edit itself is, "Let's select the entire document and rewrite completely, or do you want to have a very targeted custom behavior." And when we first launched the model, we would bias the model towards more rewrites because we saw the quality of the rewrites were much higher. But over time you are shifting based on user feedback and what you're learning from iterative deployment.

中文翻译:

以编程为例，还有一个问题是：模型在完全重写文档和进行非常具体的针对性编辑之间表现如何？这是编辑行为内部的另一层决策边界：“是选择整个文档并完全重写，还是进行非常精准的定制化行为？”当我们最初发布模型时，我们会让模型倾向于更多地重写，因为我们发现重写的质量要高得多。但随着时间的推移，你会根据用户反馈和从迭代部署中学到的东西进行调整。

(00:17:02) Karina Nguyen

English:

Lastly, the third behavior that we taught synthetically the model is how to make comments on any document. So the way we used that is we would use o1 model to seem a way of user conversation, let's say like, "Write me a document about XYZ." But then we used o1 to produce the document and then we injected user prompt to be like, "Oh, make some comments, critique my piece of writing or critique this piece of writing that you just made." And then we taught the model to make comments on the document on very specific [inaudible 00:17:45] So it's also what kind of comments you want the model to make. Do they make sense or not? How do you teach the quality of that? And it all came down to measuring progress via very robust evals. But, yeah, this is how you used o1 and a synthetic data generation for the training.

中文翻译:

最后，我们通过合成方式教给模型的第三个行为是如何对任何文档进行评论。我们的做法是：使用 o1 模型模拟用户对话，比如“给我写一份关于 XYZ 的文档”。然后我们让 o1 生成文档，接着注入用户提示词，比如“哦，给点评论，批评一下我的写作，或者批评一下你刚才写的这段话”。然后我们教会模型在文档的特定位置发表评论。这还涉及到你希望模型发表什么样的评论。它们是否有意义？你如何界定评论的质量？这一切最终都归结为通过非常鲁棒的评估（evals）来衡量进度。没错，这就是你如何利用 o1 和合成数据生成来进行训练的。

(00:18:07) Lenny Rachitsky

English:

Okay, that's so interesting. So you talk about this idea of teaching the model and you mentioned how it's using synthetic data to teach the model different behaviors is a simple way to think about it. Basically that's where you do that by showing it what success looks like using basically evals. Is that the simple way to think about it? Like, "Here's what you doing this successfully would look like," and that teaches it, "Okay, I see this is what I should be doing [inaudible 00:18:31]"

中文翻译:

好的，这太有意思了。你谈到了教导模型的想法，并提到使用合成数据来教导模型不同的行为，这是一种简单的理解方式。基本上，你是通过评估（evals）向它展示“成功”是什么样子的。可以这样简单理解吗？比如：“这就是你成功完成任务的样子”，然后这教会了它：“好的，我明白了，这就是我应该做的。”

(00:18:30) Karina Nguyen

English:

Yeah, great. Yeah, amazing. Yeah, you got it.

中文翻译:

是的，没错。太棒了，你理解得很到位。

(00:18:33) Lenny Rachitsky

English:

Okay, got it. I want to start unpacking what your day-to-day looks like as you're building these sort of things. Is it like you sitting there talking to some version of ChatGPT, crafting these evals?

中文翻译:

好的，明白了。我想开始拆解一下你构建这些东西时的日常生活是怎样的。是你坐在那里，和某个版本的 ChatGPT 对话，然后编写这些评估（evals）吗?

(00:19:19) Karina Nguyen

English:

Sometimes I do that. Sometimes I do sit with ChatGPT. Actually, I think I learned this so much from Anthropic, is people spend so much time prompting models and where quality's a really bad batch all the time, and you actually get a lot of new ideas of how do you make the model better? It's like, "This response is kind of weird. Why's it doing this?" And you start debugging or something, or you start figuring out new methods of how do you teach the model to respond in the different way, have better personality, let's say.

中文翻译:

有时我会那样做。有时我确实会和 ChatGPT 坐在一起。实际上，我觉得我从 Anthropic 学到了很多，那就是人们花大量时间给模型写提示词（prompting），当遇到一批质量很差的回复时，你实际上会产生很多关于如何改进模型的新想法。比如：“这个回答有点奇怪。它为什么要这么做？”然后你开始调试，或者开始寻找新方法，教模型以不同的方式回应，比如让它更有个性。

(00:19:19) Karina Nguyen

English:

So it's the same thing of how personality is made in the models with those. It's very similar methods. But, yes, I think my time at OpenAI have changed. I think when I first came, I was mostly research IC work so I was like building a lot of... I was running code, training models, write evals, working with PMs and designers to learn, teach them how to even think about evaluation. I think that was really cool experience and I think it was just like an adoption of, "How do we do this product management of AI feature for our AI models?" Yeah, but now it's mostly management and mentorship. I'm still doing IC research code up to 4:00 PM, although. But I just kind of changed.

中文翻译:

模型个性的塑造也是用类似的方法。但是，是的，我在 OpenAI 的时间分配发生了变化。刚来的时候，我主要做研究 IC（独立贡献者）的工作，所以我写很多代码、训练模型、编写评估，并与产品经理（PM）和设计师合作，教他们如何思考评估。我认为那是很酷的经历，就像是在探索“我们如何为 AI 模型进行 AI 功能的产品管理？”现在我主要负责管理和指导。不过，下午 4 点之前我仍然在写 IC 研究代码。只是角色发生了一些转变。

(00:20:21) Lenny Rachitsky

English:

All right, don't talk too much about being a manager.

中文翻译:

好吧，别聊太多当经理的事。

(00:20:23) Karina Nguyen

English:

Okay.

中文翻译:

好的。

(00:20:23) Lenny Rachitsky

English:

Because everyone's in firing their managers. "Who needs managers anymore?" That's what I hear now. Just kidding. It's interesting that so much of your time was spent on teaching product teams how evals integrate and how important it is. And I've heard this a few times and I haven't personally experienced it yet, so I think it's an important thread to follow is just how writing these evaluations is going to become increasingly an important part of the job of product teams, especially when they're building AI features and working with LLMs. So can you just talk a bit more about what that looks like? Is it sitting there with an Excel spreadsheet basically showing, "Here's the input, here's the output, here's how good the result was"? Talk about what that actually looks like very practically.

中文翻译:

因为现在大家都在裁撤经理。“谁还需要经理啊？”我现在听到的都是这种话。开个玩笑。很有意思的是，你花了这么多时间教产品团队如何整合评估（evals）以及它的重要性。我听过好几次这种说法，虽然我还没亲身经历过，但我认为这是一个很重要的思路：编写这些评估将如何日益成为产品团队工作的重要组成部分，特别是在构建 AI 功能和使用大语言模型时。你能再多谈谈那是什么样子的吗？是坐在那里用 Excel 表格，基本上显示“这是输入，这是输出，这是结果有多好”吗？谈谈实际操作中是什么样的。

(00:21:02) Karina Nguyen

English:

It certainly depends on what you're developing, but there are various types of evaluations. Sometimes I do ask product managers, or there's also new roles that we have, model designers, to go through some of the user feedback maybe or think of various user conversations that should have triggered... Under these circumstances, it should trigger Canvas. And then you have this ground truth label of, "Okay with this conversation it should look trigger Canvas, under this conversation it should not trigger Canvas." And you have this very deterministic kind of eval that for decision-making behaviors is like this.

中文翻译:

这当然取决于你在开发什么，但评估有很多种类型。有时我会要求产品经理，或者我们现在有的新角色——模型设计师，去查看一些用户反馈，或者构思各种应该触发功能的对话场景……比如在这些情况下，它应该触发 Canvas。然后你就有了一个基准真值（ground truth）标签：“好的，在这个对话下它应该触发 Canvas，在那个对话下不应该触发。”对于决策行为，你会拥有这种非常确定性（deterministic）的评估。

(00:21:46) Karina Nguyen

English:

When we were launching tasks, for example, how do you make correct schedules is actually really hard for the model. But we built out some of the deterministic evaluations that is like, "Okay, if the user says 7:00 PM, the model should say 7:00 PM." So if you can have deterministic evals whether it's pass or fail.

And the way it works is all the... Sometimes I ask product managers to just go create a double sheet, have different tabs and what's the current behavior, what's the ideal behavior and why, and some notes.

中文翻译:

例如，当我们发布任务功能（tasks）时，如何制定正确的时间表对模型来说其实非常难。但我们建立了一些确定性的评估，比如：“好的，如果用户说晚上 7 点，模型就应该说晚上 7 点。”这样你就能拥有非黑即白的确定性评估。具体做法是……有时我会让产品经理去创建一个表格，设置不同的标签页，记录当前行为是什么、理想行为是什么、为什么，以及一些备注。

(00:22:27) Karina Nguyen

English:

And sometimes they usually use it with evals, sometimes we use it for training. Because if you give the spreadsheet to o1 model, it can probably figure out how to teach itself a good behavior. And I think there are second type of evals that is more prevalent is human evaluations. And you can have specific trainers or you can have internal people to when you have a conversation of the prompt and then you have various completion of models, you choose the win rate. Which model is the best? Which model produce the highest quality comment or edit? And then you can have continuous win rates. And as you develop new models it should always win over the previous models. So it depends on what you want to measure.

中文翻译:

有时他们将这些用于评估，有时我们将其用于训练。因为如果你把这个电子表格给 o1 模型，它可能就能弄清楚如何教自己表现出良好的行为。我认为第二种更普遍的评估类型是人工评估（human evaluations）。你可以有专门的训练员，或者让内部人员参与：当你有一个提示词对话，然后有多个模型的完成结果时，你选择胜率（win rate）。哪个模型最好？哪个模型生成的评论或编辑质量最高？然后你可以持续跟踪胜率。当你开发新模型时，它应该始终胜过之前的模型。所以这取决于你想衡量什么。

(00:23:22) Lenny Rachitsky

English:

So interesting. Basically what I'm hearing, and there's something I'm learning about as I talk to people, is product development might move from this, "Here's a spec PRD, let's build it together and then cool, let's review it. Are we happy with this?" From that to, "Hey, AI, build this thing for me and here's what correct looks like," and I'm spending all my time on what does correct look like on evals essentially.

中文翻译:

太有意思了。基本上我听到的是——这也是我通过与人交谈学到的——产品开发可能会从“这是规格说明书 PRD，我们一起来构建它，然后评审，看我们是否满意”这种模式，转变为“嘿，AI，帮我构建这个东西，这是‘正确’的标准”，而我把所有时间都花在通过评估（evals）来定义什么是“正确”上。

(00:23:47) Karina Nguyen

English:

You definitely want to measure progress of your model and this is where evals is, is because you can have prompted model as a baseline already. And the most robust evals is the one where prompted baselines get the lowest score or something. And then because then you know if you're trained a good model, then it should just hill climb on that eval all the time, while not also regressing on other intelligence evals.

That's what I'm saying, it's more of an art than science. It's like, "Okay, if you optimize the model for this behavior, you don't want to brain damage in other areas of intelligence or..." This is happening all the time in every lab, in every research team.

中文翻译:

你肯定想衡量模型的进度，这就是评估的作用。因为你可以把仅靠提示词（prompted）的模型作为基准。最鲁棒的评估是那种让提示词基准得分最低的评估。因为这样你就知道，如果你训练出了一个好模型，它就应该在那个评估上不断“爬坡”（提升），同时又不会在其他智能评估上退步。这就是我说的，这更像是一门艺术。比如：“好吧，如果你针对这种行为优化模型，你不希望它在其他智能领域变笨……”这种情况在每个实验室、每个研究团队中都在发生。

(00:24:35) Karina Nguyen

English:

I would say prompting is also a way to prototype new product ideas. Early days at Anthropic when I was working file uploads feature, I remember I was just prompting the model to just... I remember we were launching a hundred key contexts. I was just prototyping this in their local browser. I did the demo. People really, really loved it. And they just wanted API for file uploads or something. And then that's when it clicked to me, and also one of the blog posts a long time ago, it clicked on me prompting is a new way of product development or prototyping for designers and for product managers.

中文翻译:

我想说，写提示词（prompting）也是一种原型化新产品想法的方式。在 Anthropic 早期，当我负责文件上传功能时，我记得我只是通过提示词让模型……我记得当时我们要发布 10 万（100K）上下文。我只是在本地浏览器中做原型。我做了演示，大家非常喜欢。他们甚至想要文件上传的 API 之类的。就在那时我突然开窍了，很久以前的一篇博文也提到过，我意识到提示词是设计师和产品经理进行产品开发或原型设计的一种新方式。

(00:25:20) Karina Nguyen

English:

For example, one of the features that I want to do is have a personalized starter prompts. So whenever you come to Claude, it should recommend you starter prompts based on what your interests are. And so you can literally do it prompting for that.

中文翻译:

例如，我想做的一个功能是提供个性化的初始提示词。这样每当你打开 Claude 时，它都会根据你的兴趣向你推荐初始提示词。你完全可以通过写提示词来实现这个功能的原型。

(00:25:42) Lenny Rachitsky

English:

Mm-hmm. To experiment with that.

中文翻译:

嗯，用它来进行实验。

(00:25:44) Karina Nguyen

English:

Another feature was generating titles for the conversations. It's a very small micro experience but I'm really proud of. The way we did that was we took five latest conversation from the model, asked the model, "What's the style of the user?" And then for the next new conversation, the generated title will be of the same style. It's just like really little micro experiences like this.

中文翻译:

另一个功能是为对话生成标题。这是一个非常微小的体验，但我非常自豪。我们的做法是：从模型中提取最近的五次对话，问模型：“用户的风格是什么？”然后对于下一次新对话，生成的标题就会采用相同的风格。就是这种非常细微的体验。

(00:26:12) Lenny Rachitsky

English:

That's so cool. Did you do that at Anthropic or at OpenAI?

中文翻译:

太酷了。你是在 Anthropic 还是在 OpenAI 做的这个?

(00:26:14) Karina Nguyen

English:

At Anthropic.

中文翻译:

在 Anthropic。

(00:26:16) Lenny Rachitsky

English:

Okay, cool. I love the file upload feature that Claude has by the way. ChatGPT doesn't have that yet, is that right?

中文翻译:

好的，酷。顺便说一下，我非常喜欢 Claude 的文件上传功能。ChatGPT 还没有那个功能，对吧?

(00:26:16) Karina Nguyen

English:

I think has the way.

中文翻译:

我想它有类似的方式。

(00:26:23) Lenny Rachitsky

English:

[inaudible 00:26:23]

中文翻译:

（听不清）

(00:26:22) Karina Nguyen

English:

I think the way it's implement is very different though.

中文翻译:

不过我觉得它的实现方式非常不同。

(00:26:25) Lenny Rachitsky

English:

Okay. Maybe it's the PDF feature, because I use it all the time with Claude.

中文翻译:

好吧。也许是 PDF 功能，因为我经常在 Claude 上用它。

(00:26:28) Karina Nguyen

English:

Yeah.

中文翻译:

是的。

(00:26:28) Lenny Rachitsky

English:

Okay.

中文翻译:

好的。

(00:26:28) Karina Nguyen

English:

That's cool.

中文翻译:

那很酷。

(00:26:29) Lenny Rachitsky

English:

Somebody needs to get on that. Main, it's wild how many features you built that I use every day and that many people use every day. This prototyping point you made is really important. It's something that comes up a ton on this podcast also of how that... is maybe the way that AI has most impacted the job of product builders recently is just prototyping instead of going from showing just like, "Here's a PRD, here's a design." PMs are more and more just, "Here's the prototype with the idea that I have," and it's working. You can play with it.

中文翻译:

得有人去跟进一下。天哪，你构建了这么多我每天都在用、很多人每天都在用的功能，这太疯狂了。你提到的关于原型的观点非常重要。在这个播客中也经常提到，AI 最近对产品构建者工作影响最大的方式可能就是原型设计——不再只是展示“这是 PRD，这是设计图”，产品经理越来越多地直接展示“这是我那个想法的原型”，而且它是可以运行、可以交互的。

(00:26:54) Karina Nguyen

English:

Yeah.

中文翻译:

是的。

(00:26:55) Lenny Rachitsky

English:

Yeah. Okay, I want to spend a little more time on how you operate. So you talked about you built this in launch of this tasks feature, is that the way to describe your tasks?

中文翻译:

好的。我想再多花点时间聊聊你是如何运作的。你谈到了你构建并发布了这个“任务”（tasks）功能，是这样描述的吗？

(00:27:04) Karina Nguyen

English:

Yeah.

中文翻译:

是的。

(00:27:06) Lenny Rachitsky

English:

So talk about how that emerged and let's better understand just how you collaborate with product teams and how OpenAI works in that way, whatever you can share there.

中文翻译:

那谈谈它是如何产生的，让我们更好地理解你是如何与产品团队协作的，以及 OpenAI 在这方面是如何运作的，只要是能分享的都可以。

(00:27:14) Karina Nguyen

English:

I think Canvas and tasks are going into the bucket of projects where it's more short or medium terms. And actually the way Canvas and tasks came about to be was it started with one person prototyping and creating a spec. It's kind like PRD. It's like creating a spec of the behavior of the model. I don't think tasks is extremely groundbreaking feature necessarily. What makes it really cool is because the models are so general... Model can now search, they can write sci-fi stories, they can search for stocks, they can summarize the news every day. Because the models are so general giving something familiar to people that notifications is very familiar, having reminders is very familiar. So feeling like a form factor for the people who are very familiar, same as Canvas, Google Docs is very familiar, but then you add magical AI moment and it becomes very powerful.

中文翻译:

我认为 Canvas 和任务功能属于中短期项目的范畴。实际上，Canvas 和任务功能的诞生始于一个人的原型设计和规格说明（spec）编写。这有点像 PRD，就是为模型的行为制定规格。我不认为任务功能本身一定是极其开创性的。它的酷点在于模型现在非常通用……模型现在可以搜索、写科幻故事、查股票、总结每日新闻。正因为模型如此通用，给人们一些熟悉的东西——比如通知很熟悉，提醒也很熟悉——这种对人们来说非常熟悉的表现形式（form factor），就像 Canvas 之于 Google Docs 一样，再加上神奇的 AI 时刻，它就会变得非常强大。

(00:28:26) Karina Nguyen

English:

But the way it comes usually operationally... Yeah, size is like a prototype, literally prompted prototype of how you would want the model to behave. For tasks, for example, you need to design... Literally design thinking is like okay, well, if the user says, "Remind me to go to lunch at 8:00 AM tomorrow," what information does the model need to extract from that prompt in order to create a reminder? And so this is how you design a spec for a new feature, like a tool. Canvas and tasks are all tools. So it's like how do you create the tool stack?

中文翻译:

但在操作层面上……是的，规模就像一个原型，字面意义上是通过提示词构建的模型行为原型。以任务功能为例，你需要设计……字面意义上的设计思维就是：好吧，如果用户说“提醒我明天早上 8 点去吃午饭”，模型需要从那个提示词中提取什么信息才能创建一个提醒？这就是你为新功能（比如一个工具）设计规格的方式。Canvas 和任务功能都是工具。所以这关乎你如何创建工具栈。

(00:29:09) Karina Nguyen

English:

And then it's mostly like developing JSON schema. It was like, "Okay, from this problem maybe the model should extract the time that the user requested." And then you think about which format do you want the time to be? And then how do you want the model to notify you is basically the user should give instruction to the model. And then this instruction would fire off every day or something at that particular time. So, for example, if you say, "Every day I want to learn know about the latest AI news," the model should rewrite into, "Okay search for the latest AI news and this task will get fired at that particular type that the user requested."

中文翻译:

然后这主要就像是开发 JSON 架构（schema）。比如：“好的，从这个问题中，模型也许应该提取用户请求的时间。”然后你考虑希望时间是什么格式？接着你希望模型如何通知你，基本上就是用户应该给模型下达指令。然后这个指令会在每天的那个特定时间触发。例如，如果你说“每天我都想了解最新的 AI 新闻”，模型应该将其重写为：“好的，搜索最新的 AI 新闻，这个任务将在用户要求的那个特定时间触发。”

(00:30:02) Karina Nguyen

English:

And then your design is like tool spec. Actually, I don't know. I feel like sometimes it's through conversations I... Either people ask me to join the [inaudible 00:30:15] team and they're like, "Oh my god, we need researchers." Or like, "We need some support. We need to train the models," or sometimes. Canvas was mostly like I just pitched the idea of... It got staffed quite immediately during the break, so it's dependent on the project. And then usually with staffing is mostly a product manager, model designer, actual product designer, a couple of researchers and a bunch of applied engineers. Depends on the complexity of a project. And then for tasks it took, I don't know, like two months or so to go from zero to one basically.

中文翻译:

然后你的设计就像是工具规格。实际上，我也不知道。我觉得有时是通过对话……要么是有人邀请我加入某个团队，说：“天哪，我们需要研究员。”或者说：“我们需要一些支持，我们需要训练模型。”Canvas 主要是因为我提出了这个想法……在休息期间它很快就配齐了人员，所以这取决于项目。通常人员配置包括一名产品经理、一名模型设计师、一名真正的产品设计师、几名研究员和一群应用工程师。这取决于项目的复杂程度。对于任务功能，从零到一基本上花了大约两个月的时间。

(00:30:54) Lenny Rachitsky

English:

Oh wow.

中文翻译:

噢，哇。

(00:30:54) Karina Nguyen

English:

For Canvas this was like four, five months, I guess, to go from zero to one. And then you teach product managers how to build evals and maybe how do we not only ship the better feature, but how do we think

longer term? What kind of cool features did you want tasks to have? I think it would be nice for tasks to be a little bit more personalized. It'd be nice to have to create tasks via voice on a mobile, right? This is how you get research roadmap right here is thinking how the feature will be developed in the future.

中文翻译:

对于 Canvas，从零到一大概花了四五个月。然后你教产品经理如何建立评估，也许还有我们如何不仅发布更好的功能，而且如何进行长远思考？你希望任务功能有哪些酷炫的特性？我认为如果任务功能能更个性化一点就好了。如果能在手机上通过语音创建任务，那就太棒了，对吧？这就是你获得研究路线图（roadmap）的方式——思考该功能未来将如何发展。

(00:31:39) Karina Nguyen

English:

And then from there it's like you start getting data sets. With evals, you want to make sure that goes well. And then you need to have a trade-off between what methods you want to use. And the reason why I really love relying purely on synthetic data instead of collecting data from humans is because it's much more scalable, it's cheap, less than half. You literally sample from the model and you teach the core behaviors of the models and that will generalize to all sorts of diverse coverage.

中文翻译:

从那以后，你就开始获取数据集。通过评估，你要确保一切进展顺利。然后你需要在想使用的方法之间进行权衡。我之所以非常喜欢纯粹依赖合成数据而不是从人类那里收集数据，是因为它的可扩展性要强得多，而且很便宜，成本不到一半。你直接从模型中采样，教给模型核心行为，这些行为将推广到各种多样化的场景中。

(00:32:15) Karina Nguyen

English:

And when you launch the beta feature, you learn so much from the users that you can... All your synthetic sets can be shifted in the distribution and how the users behave on the product behavior. And this is how we improve. And this is what happened with Canvass too when we launched from beta to GA.

中文翻译:

当你发布 Beta 版功能时，你会从用户那里学到很多，你可以……根据用户在产品中的行为表现，调整所有合成数据集的分布。这就是我们改进的方式。Canvas 从 Beta 版到正式发布（GA）的过程也是如此。

(00:32:34) Lenny Rachitsky

English:

Okay. This episode is brought to you by Loom. Loom lets you your screen, your camera and your voice to share video messages easily. Record a Loom and send it out with just a link to gather feedback, add context or share an update. So now you can delete that novel link email that you were writing. Instead, you can record your screen and share your message faster. Loom can help you have fewer meetings and make the meetings that you do have much more productive.

中文翻译:

好的。本集节目由 Loom 赞助。Loom 让你能够通过屏幕、摄像头和语音轻松分享视频消息。录制一段 Loom，只需一个链接即可发送出去，用于收集反馈、增加背景信息或分享更新。所以现在你可以删掉你正在写

的长篇大论的邮件了。相反，你可以录制屏幕并更快地分享你的信息。Loom 可以帮助你减少会议，并让你参加的会议更具生产力。

(00:33:02) Lenny Rachitsky

English:

Meetings start with everyone on the same page and end early. Problem solved, time saved. We know that everyone isn't a one-take wonder when it comes to recording videos. So Loom comes with easy editing and AI features to help you record once and get back to the work that counts. Save time, align your team, stay connected and get more done with Loom. Now part of Atlassian, the makers of Jira. Try Loom for free today at Loom.com/Lenny. That's L-O-O-M.com/Lenny.

中文翻译:

会议开始时大家达成共识，并提前结束。问题解决了，时间节省了。我们知道，并不是每个人在录制视频时都能“一遍过”。所以 Loom 提供了简单的编辑和 AI 功能，帮助你录制一次就能回到重要的工作中。使用 Loom 节省时间、对齐团队、保持联系并完成更多工作。Loom 现在是 Jira 制造商 Atlassian 的一部分。今天就在 Loom.com/Lenny 免费试用 Loom。网址是 L-O-O-M.com/Lenny。

(00:33:34) Lenny Rachitsky

English:

Something that I want to help people understand, and I don't even 100% understand this, is what's the simplest way to understand the job of a researcher versus say a model designer and other folks involved? What's the simplest way to understand what researchers do at OpenAI?

中文翻译:

我想帮助大家理解的一件事——甚至我自己也不是 100% 理解——就是理解研究员（researcher）与模型设计师（model designer）以及其他参与人员的工作区别，最简单的方法是什么？在 OpenAI，研究员的工作最简单的理解方式是什么？

(00:33:48) Karina Nguyen

English:

So the project that I described are mostly product-oriented. Research is mostly product research. Another component of my team is actually more longer term exploratory projects. And it's more about developing new methods, understanding those methods under a variety of circumstances. So basically developing methods, you need to follow very similar recipe of building evals but it's much more sophisticated evals. You want to have outer distribution or if you want to measure generalization, you need to capture that.

中文翻译:

我刚才描述的项目大多是面向产品的。研究主要是产品研究。我团队的另一个组成部分实际上是更长期的探索性项目。它更多是关于开发新方法，并在各种情况下理解这些方法。所以基本上开发方法时，你需要遵循与建立评估非常相似的配方，但那是更复杂的评估。你想要处理分布外（out-of-distribution）的情况，或者如果你想衡量泛化能力（generalization），你需要捕捉到这一点。

(00:34:26) Karina Nguyen

English:

But it is basically more sciencey in a way where... If we talk about synthetic data, one of the hardest things about synthetic data is how do you make it more diverse? Diversity in synthetic data is one of the most important questions right now. And so it's like exploring ways to inject diversity as a general method that will work for all is one of the research explorations. Other ones is more developing new capabilities. I feel like it's always about you work on this new method and you have signs of life that it's working, either you think of how do you make it more general or you think of how do you make it very useful? And this is how the longer-term projects become more medium, short-term project.

中文翻译:

但在某种程度上，它更具科学性……如果我们谈论合成数据，合成数据最难的事情之一就是：如何让它更多样化？合成数据的多样性是目前最重要的课题之一。所以，探索如何将多样性作为一种通用的方法注入，使其适用于所有场景，就是研究探索之一。其他的则是开发新能力。我觉得这总是关乎：你研究这种新方法，当你看到它奏效的迹象时，你要么考虑如何让它更通用，要么考虑如何让它非常有用。这就是长期项目如何变成中短期项目的方式。

(00:35:15) Lenny Rachitsky

English:

That makes sense. Essentially working on developing ways to make the model smarter, o4, o5, o6. New ways to... o1 was a big breakthrough, right?

中文翻译:

明白了。本质上是致力于开发让模型更聪明的方法，比如 o4、o5、o6。寻找新方法……o1 是一个重大突破，对吧?

(00:35:25) Karina Nguyen

English:

Yeah.

中文翻译:

是的。

(00:35:25) Lenny Rachitsky

English:

The way it operates where it's not just, "Here's your answer," it actually thinks and takes time to think through the process of coming up with an answer. Okay.

中文翻译:

它的运作方式不再仅仅是“这是你的答案”，它实际上会思考，并花时间思考得出答案的过程。好的。

(00:35:33) Karina Nguyen

English:

Yeah.

中文翻译:

是的。

(00:35:34) Lenny Rachitsky

English:

Very helpful. Speaking of that, of thinking about the future, where things are going, I want to spend some time on just this insight that basically you are building the cutting edge of AI, at the very bleeding edge of where AI is going and where it is. And so I'm very curious to hear just your take on how you think things are going to change in the world and how people work based on where you see things are going. And I know it's a broad question, but let's say in the next three years, how do you see the world changing? How do you see people's way of working changing?

中文翻译:

非常有帮助。说到思考未来和事物的发展方向，我想花点时间谈谈这样一个见解：你基本上是在构建 AI 的最前沿，处于 AI 现状和未来走向的最尖端。所以我非常好奇，基于你所看到的趋势，你认为世界会发生怎样的变化，人们的工作方式会如何改变？我知道这是一个很宽泛的问题，但假设在未来三年内，你认为世界会如何改变？人们的工作方式会如何改变？

(00:36:08) Karina Nguyen

English:

It's a very humbling experience to be in both labs, I guess. To me when I first came to Anthropic and I was like, "Oh no, I really love front-end engineering." And then the reason why I switched to research is because I realized at that time it's like, "Oh my god, Claude is getting better at front-end. Claude is getting better at coding. I think Claude can develop new apps or something and so it can develop new features for the thing that I'm working." So it was kind like this meta realization where it's like, "Oh my god, the world is actually changing." And when we first launched 100K context at that time, obviously I'm thinking about form factors that's like file uploads were very natural, very familiar to people. But you can imagine we could just make infinite chats in the Claude.ai app, as if it's 100K context.

中文翻译:

能在两家实验室工作是一种非常令人谦卑的经历。对我来说，当我刚加入 Anthropic 时，我觉得：“噢不，我真的很喜欢前端工程。”后来我转向研究的原因是，我当时意识到：“天哪，Claude 的前端能力越来越强了。Claude 的编程能力越来越强了。我觉得 Claude 甚至可以开发新的应用程序，或者为我正在做的东西开发新功能。”所以这就像是一种“元认知”层面的感悟：“天哪，世界真的在改变。”当我们第一次发布 10 万上下文时，我显然在思考表现形式（form factors），比如文件上传对人们来说非常自然、非常熟悉。但你可以想象，我们也可以直接在 Claude.ai 应用中制作无限长的聊天，就像它是 10 万上下文一样。

(00:37:04) Karina Nguyen

English:

But because file uploads... It's like form follows function. It's like the form factor, the file uploads can enable people to just literally upload anything, the books, any reports, financial and ask any task to the model. And then I remember it was either enterprise customers, financial customers were really

interested in that. It's like, "Oh wow." It's actually one of the very common tasks that people do in that setting. It's kind crazy to see how some of the redundant tasks are getting automated basically by these smart models.

中文翻译:

但因为文件上传……这就像是“形式追随功能”。文件上传这种形式可以让人们直接上传任何东西：书籍、任何报告、财务报表，并向模型布置任何任务。我记得当时无论是企业客户还是金融客户都对此非常感兴趣。就像是：“噢，哇。”这实际上是人们在那种环境下非常常见的任务之一。看到一些重复性的任务基本上被这些智能模型自动化了，感觉挺疯狂的。

(00:37:48) Karina Nguyen

English:

And they're entering the era where, I actually don't know for example sometimes if o1 gives me the correct answer or not because I'm not an expert in that field. And it's like, "I don't even know how to verify the outputs of the models." It's because all my experts know they can verify this. So, yes, so basically there are trends that are going on. The first trend is the cost of reasoning and intelligence is drastically going down.

中文翻译:

我们正在进入这样一个时代：例如，有时我实际上不知道 o1 给我的答案是否正确，因为我不是那个领域的专家。就像是：“我甚至不知道如何验证模型的输出。”因为只有专家知道如何验证。所以，基本上有一些趋势正在发生。第一个趋势是推理和智能的成本正在大幅下降。

(00:38:22) Karina Nguyen

English:

I had a blog post about this. Maybe I should update on latest benchmarks, because at that time everybody was doing one benchmark and they'd be... quickly saturated the benchmarks. So I'm like, "Now we need to do the same plot but with another frontier eval." But the cost of intelligence is going down because it becomes that much cheaper. Small models are becoming even smarter than large models and that's because of the distillation research.

中文翻译:

我曾写过一篇关于此的博文。也许我应该根据最新的基准测试进行更新，因为当时大家都在做一个基准测试，而且很快就达到了饱和。所以我心想：“现在我们需要用另一个前沿评估来画同样的图表。”但智能的成本正在下降，因为它变得便宜得多。由于蒸馏（distillation）研究，小模型甚至变得比大模型更聪明。

(00:38:56) Karina Nguyen

English:

This happened with Claude 3 Haiku. I was working with the training on the Claude 3 Haiku and I realized it was much smarter than Claude 2, which was way bigger, lots [inaudible 00:39:08]. But the power of small models become very intelligent and fast and cheap. We are moving towards that world. That has multiple implications, but the news is that people will have more access AI and that's really good. Builders and developers will have much better access to AI, but also it means all the work that has been bottlenecked by intelligence will be unblocked.

中文翻译:

Claude 3 Haiku 就是这样。我当时参与了 Claude 3 Haiku 的训练，我意识到它比 Claude 2 聪明得多，而 Claude 2 要大得多。小模型的力量在于变得非常智能、快速且廉价。我们正朝着那个世界迈进。这有多重影响，但好消息是人们将有更多机会接触 AI，这非常好。构建者和开发者将能更好地利用 AI，但也意味着所有曾受限于智能瓶颈的工作都将被打通。

(00:39:40) Karina Nguyen

English:

I'm thinking about healthcare, right? Instead of going to a doctor, I can ask ChatGPT or give ChatGPT a list of symptoms and ask me, "Would I have a cold, flu, something else?" I can literally get the access to doctor almost. And there's been some research studies around that.

中文翻译:

我在想医疗保健，对吧？与其去看医生，我可以问 ChatGPT，或者给 ChatGPT 一份症状清单，问它：“我是感冒了、流感了，还是别的什么？”我几乎可以直接获得医生的诊断。已经有一些相关的研究研究了。

(00:40:05) Lenny Rachitsky

English:

There was a New York Times story about that where they compared doctors to doctors using ChatGPT to just ChatGPT and just ChatGPT was the best of them. All doctors made it worse.

中文翻译:

《纽约时报》曾报道过一个故事，他们对比了医生、使用 ChatGPT 的医生以及纯 ChatGPT，结果纯 ChatGPT 是表现最好的。所有的医生加入后反而让结果变差了。

(00:40:18) Karina Nguyen

English:

Yeah, that's crazy. Yeah. Yeah, that's crazy, right? Education I think I would have dreamt if I had the tool like ChatGPT when I was young and would learn so much. But it's like people can now learn almost anything from these models. So they can learn new language, they can learn how to build new look apps and write anything they do want. It's humbling to have... launch Canvas and bring that thing to the people, enable them to do something else that they couldn't have ever before. There's something magical around this experience.

中文翻译:

是的，那太疯狂了。教育方面，我想如果我年轻时有 ChatGPT 这样的工具，我一定会梦寐以求，能学到太多东西。现在人们几乎可以从这些模型中学到任何东西。他们可以学习新语言，学习如何构建新应用，写任何他们想写的东西。发布 Canvas 并将其带给人们，让他们能够做以前从未做过的事情，这令人感到谦卑。这种体验中蕴含着某种魔力。

(00:40:57) Karina Nguyen

English:

Education will have massive implications. I guess like scientific research, I think it's the dream of any AI research is to automate AI research. It's kind of scary, I'd say, which makes me think that people management will stay. It's one of the hardest thing to... Emotional intelligence with the models, creativity in itself is one of the hardest things. So writers, I don't think people should be worried as much. I think will alleviate a lot of redundant tasks for people.

中文翻译:

教育将产生巨大的影响。还有科学研究，我想任何 AI 研究员的梦想都是实现 AI 研究的自动化。我想说这有点可怕，这让我觉得人员管理（people management）将会保留下来。这是最难的事情之一……模型的情商、创造力本身就是最难的事情之一。所以作家们，我认为不必过于担心。我认为它会为人们减轻很多重复性的任务。

(00:41:34) Lenny Rachitsky

English:

This is awesome. Okay, I want to follow this thread for sure. And it's funny that what you described as you were an engineer at Anthropic and you're like, "Okay, Claude is going to be very good at engineering. This isn't going to be a potentially career long term, so I'm going to move into research and AI is going to need me for a long time to build it, to make it smarter."

中文翻译:

太棒了。好的，我肯定要顺着这个思路聊。很有趣的是，你描述了你在 Anthropic 当工程师时的想法：“好吧，Claude 的工程能力会变得非常强。这可能不是一个长期的职业选择，所以我打算转向研究，AI 会在很长一段时间内需要我去构建它，让它变得更聪明。”

(00:41:53) Karina Nguyen

English:

I would say we still have... I think Canvas team has still have really cool front engineers that are really people who really care about interaction, design, interacting experience. I don't think models are there yet I think if... But we can get the models to this top 1% of front-ends and things for sure.

中文翻译:

我想说我们仍然有……我认为 Canvas 团队仍然有非常酷的前端工程师，他们是真正关心交互、设计和交互体验的人。我不认为模型目前已经达到了那个水平……但我们肯定可以让模型达到前端等领域前 1% 的水平。

(00:42:16) Lenny Rachitsky

English:

So what I want to move on to next along these lines is just, and this is just speculation, but what skills do you think will be most valuable going forward for product teams in particular? So folks are listening and they're like, "Okay, this is scary. What should I be building now to help me stay ahead and not be in trouble down the road?" What skills do you think are going to be more and more important to build?

中文翻译:

沿着这个思路，我接下来想聊的是——这纯属推测——你认为未来哪些技能对产品团队来说最有价值？听众们可能会想：“好吧，这挺吓人的。我现在应该培养什么技能，才能保持领先，不至于在未来陷入困境？”你认为哪些技能的培养会变得越来越重要？

(00:42:42) Karina Nguyen

English:

Yeah, I think creative thinking. You want to generate a bunch of ideas and filter through them and not just build the best product experience. Listening. You want to build something that the most general model will not replace you. And oftentimes you build something and you make it really, really good for specific set of users and actually the mode is now in your user feedback. The mode is more in whether you listen to them, whether you can rapidly iterate. The mode is in here. I don't think we are yet to... There are so many ideas, I think there's an abundance of ideas that you can work on. I wouldn't be worried. I feel like in fact I just think people in AI field are like... I wish they were a little bit more creative and connecting the dots across the print fields or something like that to develop really cool new generation and new paradigms of interactions with this AI.

中文翻译:

是的，我认为是创造性思维。你需要产生大量的想法并进行筛选，而不仅仅是构建最佳的产品体验。还有倾听。你想要构建一些最通用的模型无法取代你的东西。通常你构建了一些东西，并针对特定用户群做得非常非常好，实际上你的护城河（moat）现在就在你的用户反馈中。护城河更多在于你是否倾听他们，是否能快速迭代。护城河就在这里。我不认为我们已经……有这么多想法，我认为你可以尝试的想法多得是。我不会担心。事实上，我觉得 AI 领域的人……我希望他们能更有创意一点，能跨领域地把点连接起来，从而开发出真正酷炫的新一代 AI 交互范式。

(00:43:53) Karina Nguyen

English:

I don't think we've cracked this problem at all. A couple of years ago I was telling some people, I was like, "You want to build for the future." So it's like it doesn't necessarily matter whether the model is good or not, good right now, but you can build product ideas such that by the time the models will be really good, it'll work really well. I think it just happened naturally. For example, at Anthropic the Claude artifacts... And I feel early days of Canvas was, back in 2022 before ChatGPT, writing ideas was our knowledge [inaudible 00:44:36]. But I feel like Claude 1.3 model itself was not there to have made really extreme good high quality edits. For example, like coding.

中文翻译:

我认为我们根本还没有解决这个问题。几年前我告诉一些人：“你要为未来而构建。”所以，模型现在好不好并不一定重要，但你可以构建产品理念，使得当模型变得真正强大时，它能运行得非常好。我认为这是自然发生的。例如，Anthropic 的 Claude Artifacts……我觉得 Canvas 的早期阶段，早在 2022 年 ChatGPT 出现之前，写作理念就是我们的知识储备。但我觉得当时的 Claude 1.3 模型本身还不足以进行极其高质量的编辑，比如编程。

(00:44:47) Karina Nguyen

English:

And I feel like I see startups like Kaeser was doing super well. And that's because they iterate so fast. They invent new ways of training models. They move really fast. They listen to what users like, massive distributions. Yeah, it's kind of cool.

中文翻译:

我看到像 Cursor 这样的初创公司做得非常好。那是因为他们迭代得太快了。他们发明了训练模型的新方法。他们行动迅速。他们倾听用户的喜好，拥有巨大的分发量。是的，这挺酷的。

(00:45:08) Lenny Rachitsky

English:

That's really helpful actually. So what I'm hearing is that soft skills essentially are going to be more and more important, powerful. You just talked about management, leading people, being creative and coming up with innovative insights, listening. There's a post I wrote that I'll link to where I try to analyze how AI will impact product management. And we're actually very aligned, and my sense was the same thing, that soft skills are going to become more and more important. And the things that are going to be replaced is the hard skills, which is interesting because usually people value the hard skills like coding, design, writing really well. And it's interesting that AI is actually really good at that because it's taking a bunch of data, synthesizing it and writing, creating a thing, versus all these fuzzy things around of what influences, convinces people to do things and aligning and listening, like you said, creativity, anything along those lines come up as I say that.

中文翻译:

这实际上非常有帮助。所以我听到的是，软技能本质上将变得越来越重要、越强大。你刚才谈到了管理、领导他人、富有创意并提出创新见解、倾听。我写过一篇文章（我会放个链接），试图分析 AI 将如何影响产品管理。我们的观点其实非常一致，我的感觉也是一样：软技能将变得越来越重要。而将被取代的是硬技能，这很有意思，因为通常人们看重的是硬技能，比如编程、设计、优秀的写作。有趣的是，AI 实际上非常擅长这些，因为它能提取大量数据、进行综合并写作、创造东西；而相比之下，那些模糊的东西——比如如何影响、说服人们做事，以及对齐和倾听，正如你所说的创造力——当我这么说时，你有什么想法吗？

(00:46:01) Karina Nguyen

English:

I think it's actually a really, really hard to teach the model how to be aesthetic or do really good visual design or how to be extremely creative in the way they write. I still think ChatGPT kind of sucks at writing and that's because it's bottlenecked by this creative reasoning. I think characterization is one of the most important... I think for a manager, I feel like...

中文翻译:

我认为要教会模型如何具备审美、如何做真正优秀的视觉设计，或者如何在写作方式上极具创意，实际上是非常非常难的。我仍然觉得 ChatGPT 在写作方面挺烂的，那是因为它受限于创造性推理。我认为性格刻画是最重要的之一……我认为对于一个经理来说，我觉得……

(00:46:28) Karina Nguyen

English:

Actually, AI research progress is bottlenecked by management, research management. It's because you have constrained set of compute and you need to allocate the compute to the research paths that you feel the most convinced about. It was like you need to have a really high conviction in the research paths to put the compute, and it's more return on investment kind of situation. It's like, "Okay, I'm thinking a lot about across all my projects, which projects are higher priority?" Prioritization and also on the lower level, "Which experiments are really important to run right now and which are not?" and cut through the

line. So I was thinking prioritization, communication, management. People skills like empathy, understanding people, collaboration.

中文翻译:

实际上，AI 研究的进展正受到管理（即研究管理）的瓶颈限制。因为你的算力资源是有限的，你需要将算力分配给你最笃信的研究路径。你必须对研究路径有极高的信心才能投入算力，这更像是一种投资回报率的情况。就像是：“好吧，我在思考我所有的项目，哪些项目的优先级更高？”优先级排序，以及在更微观的层面上，“哪些实验现在运行非常重要，哪些不重要？”并果断取舍。所以我想到的是优先级排序、沟通、管理。还有人际交往能力，比如同理心、理解他人、协作。

(00:47:23) Karina Nguyen

English:

I think Canvas wouldn't be an amazing launch if it wasn't about people and I think it's a wonderful group of people. And I get a chance to work with people like Lee Byron who's a co-creator at GraphQL and some of the best Apple designers. It's so cool to see... and how do you create this collaboration between people. It's just something that's still humane, I think.

中文翻译:

我认为如果不是因为人，Canvas 就不会是一个如此精彩的发布，我认为那是一群非常棒的人。我有机会与像 Lee Byron（GraphQL 的共同创作者）以及一些最优秀的 Apple 设计师一起工作。看到人们之间如何建立这种协作是非常酷的。我认为这仍然是属于人性的东西。

(00:47:52) Lenny Rachitsky

English:

Let me just follow through a little bit. I imagine people listening are like, "Okay, but once we have AGI or SGI it's like it'll do all this." There's a world where like, "Why isn't all this done?" I think it's easy to just assume all that. I'm curious this idea of creativity and listening, why you think AI isn't good at it, other than it's just very hard to train it to do this well. Is there anything there of just why this is especially difficult for AI and LLMs to get good at?

中文翻译:

让我再深入追问一下。我猜听众可能会想：“好吧，但一旦我们有了通用人工智能（AGI）或超人工智能（SGI），它就能做这一切了。”会有一个世界问：“为什么这一切还没完成？”我觉得很容易产生这种假设。我很好奇关于创造力和倾听的想法，除了训练起来非常困难之外，你认为为什么 AI 不擅长这些？有没有什么深层原因，解释为什么 AI 和大语言模型特别难以掌握这些？

(00:48:20) Karina Nguyen

English:

I think currently it's difficult for many reasons. I think it's still an active research area and it's something that I think my team is working on. It's like, "Okay, how do we teach the models to be more creative in the writing?" And so I'm thinking this new paradigm of wise that the models think more should actually lead to better writing in itself. But when it comes down to idea generation or discriminating of what is a good visual design or not, I feel like it hasn't had learned examples from people to discriminate it very well. I do

think it's because there are not that many people who are actually really... It's not accessible to models to learn from these people I guess. So I definitely think that's why it sucks.

中文翻译:

我认为目前困难的原因有很多。这仍然是一个活跃的研究领域，也是我团队正在努力的方向。比如：“好吧，我们如何教会模型在写作中更有创意？”所以我认为，这种让模型思考更多的新范式，本身就应该带来更好的写作。但当涉及到创意生成，或者辨别什么是好的视觉设计时，我觉得它还没有从人类那里学到足够的例子来很好地进行辨别。我确实认为这是因为真正具备这种能力的人并不多……我想模型无法接触到这些人的经验来学习。所以我绝对认为这就是它目前表现不佳的原因。

(00:49:19) Lenny Rachitsky

English:

Yeah, that makes sense. Basically there's not enough of you yet, researchers teaching it to do these things, slash people that have incredible taste and creativity that can teach these things. You could argue this will come.

中文翻译:

是的，有道理。基本上是像你这样的研究员还不够多，无法教它做这些事，或者说拥有极佳品味和创造力、能教它这些事的人还不够多。你可以争辩说这些迟早会实现的。

(00:49:31) Karina Nguyen

English:

Right.

中文翻译:

对。

(00:49:31) Lenny Rachitsky

English:

But we don't need to keep going down that thread. Let me ask you a specific question. In this post I wrote, I made this argument that a lot of people disagreed with that strategy is something that AI tooling will become increasingly great at and take over. There's the sense that that's the thing that people will continue to be much better at and you can't offload to AI basically developing your strategy, telling you what to do to win. My case is, "Isn't strategy, just take all the inputs, all the data you have available, understand the world around you and come up with a plan to win?" It feels like AI and LLM would be incredibly smart at this. What's your take?

中文翻译:

但我们不需要一直纠结于这个话题。让我问你一个具体的问题。在我写的那篇文章中，我提出了一个很多人不同意的观点：AI 工具在制定“战略”方面会变得越来越出色并最终接管。人们普遍认为，战略是人类将继续保持优势的领域，你不能把制定战略、告诉你要做什么才能赢的任务交给 AI。我的理由是：“战略不就是获取所有输入、所有可用数据，理解你周围的世界，然后制定一个获胜计划吗？”感觉 AI 和大语言模型在这方面会极其聪明。你怎么看？

(00:50:10) Karina Nguyen

English:

I think so too. I think again, you teach the model all sorts of tools and capabilities and reasoning and it's like when it comes down to... For Canvas right now, it would be very cool for the model just aggregate all the feedback from users, summarize me the top five most painful flows on user experiences. And then the model itself is very capable of thinking of knowing how it's been made, figure out how to create a dataset for itself to train on it. And I don't think that we are far away from that self-improvement, models becoming self-improved by...

中文翻译:

我也这么认为。我认为，当你教给模型各种工具、能力和推理时，当涉及到……以现在的 Canvas 为例，如果模型能汇总用户的所有反馈，为我总结出用户体验中最痛苦的前五个流程，那就太酷了。然后模型本身非常有能力思考它是如何被制造出来的，并弄清楚如何为自己创建一个数据集来进行训练。我不认为我们离那种自我改进、模型通过……实现自我提升还有很远。

(00:50:54) Karina Nguyen

English:

That, and the part of development, is basically self-improving. It's kind of like its own organism or something. Again, like strategies, it's more like data analysis and coming up with... I think what models are really good at is connecting the dots, I think. It's like if you have user feedback from this source, but you also have an internal dashboard with metrics and then you have other feedback or input and then it can create a plan for you, recommendations even. And I think this is one of the most common use cases for ChatGPT too, is coming up with these sort of things.

中文翻译:

那部分开发基本上就是自我改进。它有点像一个独立的有机体。再说战略，它更像是数据分析并提出……我认为模型真正擅长的是“连接点”。比如你从这个来源获得了用户反馈，但你还有一个带有指标的内部仪表盘，然后你还有其他的反馈或输入，接着它就能为你制定计划，甚至是建议。我认为这也是 ChatGPT 最常见的用例之一，就是构思这类事情。

(00:51:47) Lenny Rachitsky

English:

That makes sense essentially a human can only comprehend so much information at once and look at so much data at once to synthesize takeaways. And as you said, these context windows are huge now. Here's all the information, what's the most important thing I should do?

中文翻译:

这很有道理，本质上人类一次只能理解有限的信息，一次只能查看有限的数据来综合结论。正如你所说，现在的上下文窗口非常巨大。把所有信息都给它，问它：“我应该做的最重要的事情是什么？”

(00:51:59) Karina Nguyen

English:

Yeah, same as scientific research. Ideally the model would be able to suggest ideas, new ideas, or iterate on the experimental given the empirical results of the previous experiments like how do you come up with new ideas or the methods?

中文翻译:

是的，科学研究也是如此。理想情况下，模型应该能够根据之前实验的经验结果提出想法、新想法，或者对实验进行迭代，比如你如何提出新想法或新方法？

(00:52:18) Lenny Rachitsky

English:

Yeah. Oh, man. Okay, so just to close the loop on this conversation, this part of the thread is the skills you're suggesting people focus on building and leaning into is soft skills like creativity, managing influence, collaboration, looking for patterns. Is that generally where your mind is at?

中文翻译:

是的。天哪。好的，为了结束这部分的讨论，你建议人们重点培养和倾向的技能是软技能，比如创造力、管理影响力、协作、寻找模式。这大致就是你的想法吗？

(00:52:40) Karina Nguyen

English:

Yeah, I'm thinking a lot about how do we make organizations more effectively and I think this is mostly management, I guess. It's like how do you organize research teams or generally teams combined... Compose teams such that they will be at their maximally succeed or at the maximal performance of what can possibly... We can literally create the next generation of computers. It's just the matter of conviction and the way you manage through that. It's scaling organizations or scaling product research, I guess.

中文翻译:

是的，我经常在思考如何让组织更有效率，我想这主要关乎管理。比如你如何组织研究团队，或者通俗地说，如何组合团队……组建团队，使他们能够最大限度地成功，或者发挥出最大的性能……我们真的可以创造下一代计算机。这只是一个信念问题，以及你如何通过管理来实现它。我想这就是扩展组织规模或扩展产品研究规模。

(00:53:15) Lenny Rachitsky

English:

Yeah, I think you're basically building this thing and not efficiently doing it is limiting the potential of the human species right now.

中文翻译:

是的，我认为你基本上是在构建这个东西，如果效率不高，实际上是在限制目前人类物种的潜力。

(00:53:16) Karina Nguyen

English:

Right.

中文翻译:

对。

(00:53:26) Lenny Rachitsky

English:

It's mismanagement within the research team in OpenAI and Anthropic and some of these other models.

中文翻译:

那是 OpenAI、Anthropic 以及其他一些模型研究团队内部的管理问题。

(00:53:32) Karina Nguyen

English:

Yeah, it's kind of crazy to think about it.

中文翻译:

是的，想想还挺疯狂的。

(00:53:33) Lenny Rachitsky

English:

Holy moly. Okay, so speaking of Anthropic and OpenAI, you've worked at both. Very few people have worked at both companies and have seen how they operate. I'm curious just what you've noticed about the differences between these two, how they operate, how they think, how they approach stuff. What can you share along those lines?

中文翻译:

天哪。好的，说到 Anthropic 和 OpenAI，你在这两家公司都工作过。很少有人在两家公司都待过并见过它们的运作方式。我很想知道你观察到的这两者之间的区别——它们如何运作、如何思考、如何处理事情。在这方面你能分享些什么？

(00:53:48) Karina Nguyen

English:

It's more similar than different. Obviously there was a lot of... There are some differences always comes to nuances. I would say culture. I really love Anthropic and I have a lot of friends there. And I also love OpenAI and they still have a lot of friends though. So it's not about enemies. I feel like there's in AI, it's all like, "Yeah, they're competitors. There's enemies." It's actually like one big community of people doing the same thing. I would say what I've learned from Anthropic is this real care and craft towards model behavior, model craft, model training.

中文翻译:

相似之处多于不同之处。显然有很多……差异总是体现在细微之处。我会说是文化。我非常喜欢 Anthropic，我在那里有很多朋友。我也喜欢 OpenAI，我在那里也有很多朋友。所以这无关敌对。我觉得在 AI 领域，大家总觉得：“是的，他们是竞争对手，是敌人。”实际上，这就像是一个大家都在做同样事情的大社区。我想说我从 Anthropic 学到的是对模型行为、模型工艺和模型训练的真正关怀和匠心。

(00:54:32) Karina Nguyen

English:

And I've been thinking a lot about, "Okay, what makes Claude Claude and what makes ChatGPT ChatGPT?" And it's like I still have some sense of operational processes that leads to the outputs, to the model. It's the outputed model. And it's like the reason why Claude has so much more personality and is more like a librarian... I don't know. I don't know. I am visualizing Claude being like a librarian at some point, very nerdy or something. ... is because I feel like it's the reflection of the creators who are making this model. And a lot of details around the character and the personality and whether the model should follow up on this question or not.

中文翻译:

我一直在思考：“好吧，是什么让 Claude 成为 Claude，又是什么让 ChatGPT 成为 ChatGPT？”我仍然能感觉到导致模型输出的操作流程。这就是输出的模型。比如 Claude 为什么更有个性，更像一个图书馆管理员……我不知道，我脑海中 Claude 的形象有时就像个管理员，很书呆子气之类的。……这是因为我觉得它是创造这个模型的创作者的反映。关于角色、个性和模型是否应该追问这个问题，有很多细节。

(00:55:19) Karina Nguyen

English:

What's the correct ethical behavior for the model in these scenarios? A lot of crafts and curated datasets. This is where I learned that part of art, I guess, at Anthropic. I would say Anthropic is much smaller. When I joined it was, what, like 70 people? When I left it was tons of people. And obviously the culture changed so much. I really enjoyed being early days startup lives, and people knew each other as a family. But the culture shifted.

中文翻译:

在这些场景中，模型的正确伦理行为是什么？大量的匠心和精心策划的数据集。我想，这就是我在 Anthropic 学到的“艺术”部分。我想说 Anthropic 要小得多。我加入时大概只有 70 人？我离开时已经有很多人了。显然文化发生了很大变化。我非常喜欢早期的初创生活，大家像家人一样互相认识。但文化发生了转变。

(00:55:53) Karina Nguyen

English:

I would say that I learned from Anthropic that they're much better at focusing and prioritization of... Very hardcore prioritization, I guess. And they need to do it. But I think OpenAI's much more innovative and much more risk-takers in terms of product or research. Actually, in way your full-time job can be just teaching the model how to be creative writers. And it's like there's some luxury in this research freedom that comes with scale, maybe. I don't know. I'd say I have much more creative product freedom to do almost anything, I guess, within OpenAI, evolve ChatGPT into the vision that we want. It's more probably bottoms-up, I guess.

中文翻译:

我想说，我从 Anthropic 学到的是，他们更擅长专注和优先级排序……我想是那种非常硬核的优先级排序。他们必须这样做。但我认为 OpenAI 在产品或研究方面更具创新性，更敢于冒险。实际上，在某种程度上，你的全职工作可以只是教模型如何成为创意作家。这种研究自由可能伴随着规模而来的某种奢侈，也许吧。我不知道。我想说在 OpenAI 内部，我有更多的创意产品自由去做几乎任何事情，将 ChatGPT 演变成我们想要的愿景。我想这可能更多是自下而上的。

(00:56:51) Lenny Rachitsky

English:

Yeah, that's how I was thinking about it. It feels like OpenAI is more bottoms-up, distributed, people bubble up ideas, try stuff. And that leads to more products launching, I imagine more things just kind of being tried versus more of a, "Let's just make sure everything we do is awesome and great and craft and thinking deeply about every investment."

中文翻译:

是的，我也是这么想的。感觉 OpenAI 更多是自下而上的、分布式的，人们提出想法、尝试新事物。这导致了更多产品的发布，我猜会有更多东西被尝试，而不是那种“让我们确保我们做的每一件事都非常出色、有匠心，并深入思考每一项投入”的风格。

(00:57:08) Karina Nguyen

English:

Right.

中文翻译:

对。

(00:57:08) Lenny Rachitsky

English:

That's really interesting. I've never heard it described this way. Karina, we've covered so much ground. This is going to help a lot of people with so many ways of thinking about where the future's going. Before we get to our very exciting lightning round, I'm curious if there's anything else that you think might be helpful to share or get into?

中文翻译:

这真的很有意思。我从未听过有人这样描述。Karina，我们聊了很多内容。这将在思考未来走向方面给很多人带来启发。在我们进入非常精彩的闪电问答环节之前，我很好奇你是否觉得还有什么其他值得分享或探讨的内容？

(00:57:23) Karina Nguyen

English:

One of my regrets, I guess, when I was early days at Anthropic was that... I think there was some luxury of the time, because pre-ChatGPT, to actually come in with a bunch of ideas and prototype almost every

day. And I think that we did a lot of cool ideas like Claude, and Slack was actually one of the first tool-usey products. It's like Claude could operate in your workplace now. It's kind of cool because you can add Claude to summarize the thread. So maybe you have an entire conversation with someone and then you want a summary of what happened you can ask Claude, "Summarize this."

中文翻译:

我想，我在 Anthropic 早期的一个遗憾是……我认为当时有一种时间的奢侈，因为在 ChatGPT 出现之前，真的可以每天带着一堆想法去做原型。我认为我们做了很多酷炫的想法，比如 Claude，而 Slack 实际上是首批具有工具属性的产品之一。就像 Claude 现在可以在你的工作场所运行。这挺酷的，因为你可以添加 Claude 来总结线程。比如你和某人进行了一整段对话，然后你想要一个总结，你可以问 Claude：“总结一下这个。”

(00:58:07) Karina Nguyen

English:

Also, it was really fun to iterate on the model itself. It's like when you just talk to the model in Slack forever. It created some social element, it was kind like [inaudible 00:58:19] and this Discord, people learned so much about prompting and how to work with Claude. Actually, one of the features that was early tasks prototype is every Monday Claude would just summarize the entire channel. Or every Friday we'd just summarize a bunch of channels and give the news about the organization, or something.

中文翻译:

此外，在模型本身上进行迭代也很有趣。就像你在 Slack 里一直和模型聊天。它创造了一些社交元素，有点像 Discord，人们学到了很多关于提示词以及如何与 Claude 合作的知识。实际上，早期任务原型的一个功能是：每周一 Claude 会总结整个频道。或者每周五总结一堆频道，并提供关于组织的新闻之类的。

(00:58:48) Karina Nguyen

English:

And it's kind of like really cool form factor. I think thinking about form factor's a really important question in AI, especially we haven't even figured out how do we create an awesome product experience with o-series models. It's like the paradigm between synchronous real time give an answer paradigm into more asynchronous paradigm of agents working on the background. But then now the question is the agents should build trust with you, right? And trust builds over time, which is like with humans. And you start this collaboration which is why this collaboration model with you and the model is so important because you build trust and the model learns from your preferences so that it can become more personalized and it will start predicting the next action that you want to take on the computer or something. And it's more predictive, much more... We went from personal computers to personal model basically here.

中文翻译:

这是一种非常酷的表现形式。我认为思考表现形式是 AI 中一个非常重要的问题，特别是我们甚至还没有弄清楚如何利用 o 系列模型创造出色的产品体验。这就像是从“同步实时给出答案”的范式转变为“智能体在后台运行”的更多异步范式。但现在的问题是，智能体应该与你建立信任，对吧？信任是随着时间建立的，就像人与人之间一样。你开始这种协作，这就是为什么你与模型之间的协作模式如此重要，因为你建立了信任，模型从你的偏好中学习，从而变得更加个性化，它会开始预测你想在电脑上采取的下一个动作之类的。它更具预测性，更……我们基本上是从个人电脑时代迈向了个人模型时代。

(00:59:54) Lenny Rachitsky

English:

Why is it not a thing? That seems like such an obvious feature that every LLM should have as a Slack bot version of them. Is that a thing I can help you install? Or is that not a thing right now?

中文翻译:

为什么这还没实现？这看起来是一个非常显而易见的功能，每个大语言模型都应该有一个 Slack 机器人版本。那是现在我可以帮你安装的东西吗？还是说现在还没这回事？

(01:00:03) Karina Nguyen

English:

I know that Claude and Slack was sunset in 2023 or something. I think it was after ChatGPT was mostly the focus on customer use cases or enterprise use cases.

中文翻译:

我知道 Claude 在 Slack 上的功能在 2023 年左右停用了。我想是在 ChatGPT 之后，重点主要转向了客户用例或企业用例。

(01:00:17) Lenny Rachitsky

English:

Mm-hmm. Bummer.

中文翻译:

嗯，真遗憾。

(01:00:19) Karina Nguyen

English:

I think the form factor of Claude and Slack was kind of constrained a little bit when you want to talk about new features.

中文翻译:

我认为当你想要讨论新功能时，Claude 在 Slack 上的表现形式有点受限。

(01:00:28) Lenny Rachitsky

English:

Bummer. I want that.

中文翻译:

遗憾，我想要那个功能。

(01:00:30) Karina Nguyen

English:

I know that ChatGPT had Slackbar tools. I don't know, maybe it will come back sometime.

中文翻译:

我知道 ChatGPT 曾有过 Slack 栏工具。我不知道，也许以后会回来。

(01:00:35) Lenny Rachitsky

English:

All right, I would pay for that. Any other memories from that time of early days? Because that's a really special place to have been is early days Anthropic. Any other memories or stories from that time that might be interesting to share?

中文翻译:

好吧，我愿意为此付费。关于早期的那段时光，还有什么其他回忆吗？因为早期的 Anthropic 确实是一个非常特别的地方。那个时期还有什么有趣的回忆或故事可以分享吗？

(01:00:48) Karina Nguyen

English:

I think the very first launch when we felt... When click from use, again, was 100K context launch is when the models could input the entire book and give you a summary of the book or something. Or the financial... or catalog multi files financial reports and then give you an answer to the question, to very specific questions. I think there was something in there that was kind like, "Oh my god, this is a really cool new capability." Not model capability, but more like the capabilities that came from the product form factor itself rather than the model capability as much.

中文翻译:

我认为第一次让我们觉得……再次产生共鸣的发布是 10 万上下文的发布，当时模型可以输入整本书并为你提供总结。或者是财务……或者编目多个文件的财务报告，然后回答非常具体的问题。我认为那里面的某些东西让人觉得：“天哪，这真是一个非常酷的新能力。”这不完全是模型的能力，更多是来自产品表现形式本身的能力，而不仅仅是模型能力。

(01:01:34) Karina Nguyen

English:

I think other prototypes that we were thinking about... There's one part having a Claude workspaces and it's kind of the same idea of Claude and I would have this shared workspace and that share workspace is like a document and we can iterate on the document. And I feel like sometimes the ideas, [inaudible 01:01:55] and they're locked for two years, just like in this case.

中文翻译:

我想我们当时考虑的其他原型……有一个是 Claude 工作区（workspaces），这有点像 Claude 和我拥有一个共享工作区，那个共享工作区就像一个文档，我们可以对文档进行迭代。我觉得有时这些想法会被锁定两年，就像现在这种情况。

(01:02:00) Lenny Rachitsky

English:

It's interesting, there's these milestones that kind of open up our view of what is happening and where things are going. ChatGPT think was the first of just like, "Wow, this is much better than I would've thought." You talked about 100K context windows where you could upload a book and ask it questions and have it summarize. I actually use that all the time. When I have interview guests and they wrote a book, I sometimes don't have time to read the whole book. So I use it to help me understand what the most interesting parts are. And then I actually dive into the book, just to be clear. And then, I don't know, maybe voice was another one where you could talk to say ChatGPT. Is there any other moments there that you're like, "Wow, this is much better than I thought it was going to be?"

中文翻译:

很有意思，这些里程碑开启了我们对现状和未来走向的视野。ChatGPT 应该是第一个让人觉得“哇，这比我想象的好得多”的东西。你提到了 10 万上下文窗口，可以上传一本书并提问、总结。我实际上一直在用这个功能。当我有采访嘉宾写了书，我有时没时间读完整本书，所以我用它来帮我理解最有趣的部分。澄清一下，然后我还是会去读那本书。然后，我不知道，也许语音是另一个，你可以和 ChatGPT 说话。还有没有其他时刻让你觉得：“哇，这比我想象的要好得多？”

(01:02:39) Karina Nguyen

English:

Yeah, I think the computer use agents, like the model operating the desktop. And you can essentially think of new kind of experience where the model can learn the way you browse. And from that preference it can just browse as just like you. It's kind of simulated persona. And it's actually very similar to the idea of like, "Okay, maybe Sam Altman doesn't have a lot of time. Maybe I want to talk to his simulation and ask..." Or, for example, I really appreciate some of the technical mentorship. Yeah, cool. But he doesn't have a lot of time so it's like I really want to ask him this questions. How do you respond with simulated environments like this would be really cool.

中文翻译:

是的，我认为是“计算机使用智能体”（computer use agents），比如模型操作桌面。你基本上可以想象一种全新的体验，模型可以学习你浏览网页的方式。根据这种偏好，它可以像你一样浏览。这有点像模拟人格。这实际上与这种想法非常相似：“好吧，也许 Sam Altman 没有很多时间，也许我想和他的模拟人格交谈并提问……”或者，例如，我非常感激一些技术指导。是的，很酷。但他没有很多时间，所以我真的很想问他这些问题。在这样的模拟环境中如何回应，那会非常酷。

(01:03:37) Lenny Rachitsky

English:

That's a great place to plug Lennybot, have one of those. It's trained on all of my podcasts and newsletters.

中文翻译:

这正是宣传 Lennybot 的好时机，我就有一个。它是基于我所有的播客和新闻通讯训练的。

(01:03:42) Karina Nguyen

English:

Oh, cool.

中文翻译:

噢，酷。

(01:03:43) Lenny Rachitsky

English:

It sits on many models. I don't know which exactly they use, but it's exactly that. And it's not even me, it's all the guests that have been on the podcast and on newsletter as I wrote. And you could just ask it, "How do I grow my product? How do I develop a strategy?" And it's actually shockingly good.

中文翻译:

它运行在多个模型之上。我不知道他们具体用的是哪一个，但它正是你说的。它甚至不只是我，它包含了播客的所有嘉宾和我写过的新闻通讯。你可以直接问它：“我如何增长我的产品？我如何制定战略？”它实际上好得令人震惊。

(01:03:58) Karina Nguyen

English:

Do you feel like it reflects who you are?

中文翻译:

你觉得它反映了你是谁吗？

(01:03:58) Lenny Rachitsky

English:

Yeah.

中文翻译:

是的。

(01:03:58) Karina Nguyen

English:

Or would it be... Okay.

中文翻译:

或者它会……好的。

(01:04:01) Lenny Rachitsky

English:

The best part of it is you can talk to it. There's an ElevenLabs voice version that's trained on my voice from this podcast, and it's actually very good and people have told me they sit there for hours talking to it.

中文翻译:

最棒的部分是你可以和它说话。有一个 ElevenLabs 的语音版本，是用我这个播客的声音训练的，效果非常好，有人告诉我他们坐在那里和它聊了好几个小时。

(01:04:15) Karina Nguyen

English:

Wow.

中文翻译:

哇。

(01:04:15) Lenny Rachitsky

English:

And somebody told it, "Interview me like I am on Lenny's podcast, ask me questions about my career." And he did a half hour podcast episode with Lennybot.

中文翻译:

还有人告诉它：“像在 Lenny 的播客上一样采访我，问我关于职业生涯的问题。”然后他用 Lennybot 做了一期半小时的播客节目。

(01:04:24) Karina Nguyen

English:

Oh my god, that's so fun.

中文翻译:

天哪，那太有趣了。

(01:04:27) Lenny Rachitsky

English:

It's incredible. Future is wild.

中文翻译:

不可思议。未来太疯狂了。

(01:04:29) Karina Nguyen

English:

Yeah. I think content transformation is... I would imagine sometime when you generate a sci-fi story in Canvas, you can transform this into audiobook where you have very natural content transformation of one media to another media. I think one of my earliest inspiration is one of the last episodes of Westworld where, I don't want to spoil, but where Dolores comes to her work at that time and she comes to this new

workspace and she starts writing a story. And then as she writes a story, a 3D, virtual reality, starts creating on the fly. So I kind want to create that. Kind cool.

中文翻译:

是的。我认为内容转换是……我可以想象，某天当你在 Canvas 中生成一个科幻故事时，你可以将其转换为有声书，实现一种媒体到另一种媒体的非常自然的内容转换。我最早的灵感之一来自《西部世界》的最后几集（我不想剧透），Dolores 当时去上班，她来到这个新的工作空间并开始写故事。当她写故事时，一个 3D 虚拟现实就开始即时创建。我有点想创造那样的东西。挺酷的。

(01:05:24) Lenny Rachitsky

English:

Wow. Speaking of medium, I guess I was wondering if I should go in this direct or not, but real quick. Kevin Weill/Kevin Weill, I don't know exactly how to pronounce his last name, the CPO of OpenAI.

中文翻译:

哇。说到媒体，我在想是否应该往这个方向聊，但很快提一下。Kevin Weill，我不知道他姓氏的确切发音，OpenAI 的首席产品官（CPO）。

(01:05:35) Karina Nguyen

English:

Kevin Weill, uh-huh.

中文翻译:

Kevin Weill，嗯。

(01:05:37) Lenny Rachitsky

English:

Is it Weill or Weill?

中文翻译:

是发 Weill 还是 Weill?

(01:05:37) Karina Nguyen

English:

I think Weill.

中文翻译:

我想是 Weill。

(01:05:39) Lenny Rachitsky

English:

Weill. Okay. Okay. Let's just say that. We'll go with that.

中文翻译:

Weill。好的。我们就这么叫吧。

(01:05:40) Karina Nguyen

English:

I hope, yeah.

中文翻译:

希望是对的，是的。

(01:05:43) Lenny Rachitsky

English:

He did a panel at the Lenny and Friends Summit last year and he made this really fascinating point that chat is a really interesting interface for these tools because they're just getting smarter and smarter and smarter and smarter and smarter. And chat continues to work as a paradigm to just interact with them, similar to a human. You could talk to Albert Einstein. You could talk to someone not very smart and it's all conversation still. And so it's a really flexible way to interact with increasingly good intelligence. At some point it'll not be so great, and you were talking about all these ways that you're adding additional ways to interact. But it's interesting chat proved to be a really powerful layer on top of all this stuff.

中文翻译:

他在去年的 Lenny and Friends 峰会上参加了一个小组讨论，他提出了一个非常迷人的观点：聊天是这些工具的一个非常有趣的界面，因为它们变得越来越聪明。而聊天作为一种与它们交互的范式，类似于与人交流，持续奏效。你可以和阿尔伯特·爱因斯坦交谈，也可以和不太聪明的人交谈，这始终都是对话。因此，这是一种与日益增强的智能进行交互的非常灵活的方式。在某些时候它可能没那么好，而你刚才谈到了你正在增加的其他交互方式。但有趣的是，聊天被证明是覆盖在所有这些东西之上的一个非常强大的层。

(01:06:22) Karina Nguyen

English:

Yeah, that's real cool. I feel like chat also has social element which is very humane. It's like, yeah, you sometimes want to get into group chat. And having conversations with AI is kind like a group chat in itself, as messaging. Actually, this idea of how do you build features like this? I see tasks as this general feature that will scale very nicely as the models would develop new capabilities themselves. The models will be able to do better searches and create new... come up with more creative writing on render, react apps and like HTML apps. And you can have everyday new puzzle for you, every day continue the story from the previous days. It scales very nicely.

中文翻译:

是的，那真的很酷。我觉得聊天也有社交元素，这非常人性化。就像，是的，你有时想加入群聊。与 AI 对话本身就像是一种群聊，一种发消息。实际上，关于如何构建这样的功能？我认为任务功能是一个通用的功能，随着模型自身开发出新能力，它会扩展得非常好。模型将能够进行更好的搜索，并创造新的……在渲染、React 应

用和 HTML 应用上提出更具创意的写作。你可以每天都有一个新的谜题，每天继续前一天的故事。它的扩展性非常好。

(01:07:14) Lenny Rachitsky

English:

You mentioned something as we were getting into this extra section that we ended up going down is this idea of the agents using a computer. I know this is actually something you are going to launch today, the day we're recording it, which will be out by the time this comes out, called Operator, can you talk about this very cool feature that people will have access to?

中文翻译:

在我们进入这个额外话题时，你提到了智能体使用计算机的想法。我知道这实际上是你们今天（我们录音的这一天）要发布的东西，当这期节目播出时它已经发布了，叫做 Operator（操作员）。你能谈谈这个人们将能使用的非常酷的功能吗？

(01:07:33) Karina Nguyen

English:

Yeah, so I unfortunately did not work on that, but I'm really, really excited about this launch. It's basically an agent that can complete the task in its own virtual computer, in its own virtual environment. You can do any literally task like order me a book on Amazon. And then ideally the model will either follow up with you which book do you want, or know you so well that it start recommending, "Oh, here is the five books that I might recommend you to buy." And then you hit, "Yeah, help me buy." And then the model goes off into its own virtual little browser and complete the task and buy the book on the Amazon. And then if you give the model credentials, credit cards, obviously it comes with a lot of trust and safety, then it will just complete the thing for you. It's a virtual assistant.

中文翻译:

是的，遗憾的是我没有参与那个项目，但我对这次发布非常非常兴奋。它基本上是一个智能体，可以在它自己的虚拟计算机、它自己的虚拟环境中完成任务。你几乎可以布置任何任务，比如“帮我在亚马逊上订一本书”。理想情况下，模型要么会追问你想要哪本书，要么因为它非常了解你，开始推荐：“哦，这是我建议你购买的五本书。”然后你点击“是的，帮我买”。接着模型就会进入它自己的虚拟小浏览器，完成任务并在亚马逊上买书。如果你给模型凭据、信用卡（当然这涉及很多信任和安全问题），它就会为你完成这件事。它是一个虚拟助手。

(01:08:37) Lenny Rachitsky

English:

It's interesting how this just sounds like obviously this should happen. Why is this not yet a thing? Which is also mind-blowing that we're just assuming this should exist. Just some AI doing things for you on a computer we just ask it to do.

中文翻译:

很有意思，这听起来就像是理所当然应该发生的。为什么现在还没实现呢？同样令人震惊的是，我们竟然理所当然地认为这应该存在。就是让 AI 在电脑上按照我们的要求帮我们做事。

(01:08:37) Karina Nguyen

English:

Yeah.

中文翻译:

是的。

(01:08:50) Lenny Rachitsky

English:

It's absurd.

中文翻译:

这太不可思议了。

(01:08:51) Karina Nguyen

English:

It's actually really hard. And I think you're still cracking this, but feel like... I don't know if you use Tuple like a pair programming product.

中文翻译:

这实际上非常难。我认为大家仍在攻克这个难题，但感觉……我不知道你是否用过 Tuple 这样的结对编程产品。

(01:09:03) Lenny Rachitsky

English:

No.

中文翻译:

没有。

(01:09:04) Karina Nguyen

English:

But at Anthropic we loved pair programming, so if you used-

中文翻译:

但在 Anthropic 我们非常喜欢结对编程，所以如果你用过——

(01:09:09) Lenny Rachitsky

English:

Oh yeah, Shopify uses this. I remember it came up on a podcast episode.

中文翻译:

噢，是的，Shopify 在用这个。我记得在某期播客里提到过。

(01:09:10) Karina Nguyen

English:

Oh, nice. Yeah, so it is a very cool product where you can just call anyone at any time and then share screen and the other person can have access to the screen or start literally operating your computer. And it's very realtime... The allegiance is very... it's very high quality. And it's just like I kind want the same. I want to pair program with my model and the model should even talk to me. Draw very specific section in my code and just go to tell me... Obviously teach me and we can have different modes. It's like right, this is a product right here for you. I don't know. Some people should build that.

中文翻译:

噢，太好了。是的，这是一个非常酷的产品，你可以随时呼叫任何人，然后分享屏幕，对方可以访问屏幕或开始直接操作你的电脑。它是非常实时的……连接质量非常高。我想要同样的东西。我想和我的模型结对编程，模型甚至应该和我说话。在我的代码中画出非常具体的区域并告诉我……显然是教我，我们可以有不同的模式。就像是，没错，这就是为你准备的产品。我不知道，应该有人去开发这个。

(01:09:58) Lenny Rachitsky

English:

It sounds like a startup just got birthed-

中文翻译:

听起来一家初创公司刚刚诞生了——

(01:09:59) Karina Nguyen

English:

Yes.

中文翻译:

是的。

(01:10:00) Lenny Rachitsky

English:

... from someone listening to this. You mentioned that it's very hard to do this agent controlling a computer as you and helping out. What makes it so hard for whatever, however much you can explain briefly?

中文翻译:

……来自某个正在听这期节目的人。你提到让智能体像你一样控制电脑并提供帮助是非常困难的。你能简短解释一下，到底是什么让这件事如此困难吗？

(01:10:11) Karina Nguyen

English:

Much of it is because right now the model's operating on pixels instead of language or whatnot. Pixels is actually really, really hard. The models [inaudible 01:10:25] perception, or visual perception. I think there's still a lot of multimodal research that's going on, but I think language scaled so much easier compared to multimodal because of that.

中文翻译:

很大程度上是因为目前模型是在像素（pixels）上操作，而不是语言之类的。像素操作实际上非常非常难。模型在感知或视觉感知方面还不够。我认为目前仍有很多多模态研究在进行，但我认为正因如此，语言比多模态更容易扩展。

(01:10:38) Karina Nguyen

English:

Another thing that I guess my team is working that is how do you derive human intent very correctly? It's like sometimes does the model know enough information to ask a follow-up question or to complete the task? You don't want an agent to go off for 10 minutes and then come back with an answer that you didn't even want. That actually creates much more worse user experience. And this comes with teaching the model people skills. It's like, "What do people like? Kind like creating the mental model of the user and care about the user in order to ask certain questions. Actually, that part is hard to do for the models.

中文翻译:

另一件事，我想我的团队正在研究的，是如何非常准确地推导人类意图？比如，有时模型是否掌握了足够的信息来提出追问或完成任务？你不希望一个智能体消失 10 分钟，然后带回一个你根本不想要的答案。那实际上会造成更糟糕的用户体验。这涉及到教给模型社交技能。比如：“人们喜欢什么？”有点像建立用户的心理模型并关心用户，以便提出某些问题。实际上，这部分对模型来说很难做到。

(01:11:28) Lenny Rachitsky

English:

That relates to what we talked about earlier where this kind of the soft skill, people skills piece is not where these models are strong yet.

中文翻译:

这与我们之前谈到的相呼应，即软技能、社交技能这部分目前还不是这些模型的强项。

(01:11:34) Karina Nguyen

English:

Yeah.

中文翻译:

是的。

(01:11:35) Lenny Rachitsky

English:

Okay. I'm going to skip the lightning round. I want to ask just one question from the lightning round, something fun.

中文翻译:

好的。我打算跳过闪电问答环节。我只想问闪电问答中的一个问题，好玩的。

(01:11:41) Karina Nguyen

English:

Yeah.

中文翻译:

好的。

(01:11:44) Lenny Rachitsky

English:

Okay, so when AI replaces your job, Karina, I'm curious what you're... And it gives you a stipend, gives you a monthly stipend. Here's your salary for the month. What would you want to do? What do you want to spend your time on? What will you be doing in this future world?

中文翻译:

好的，Karina，当 AI 取代了你的工作，我很好奇你会……而且它还给你发津贴，每月发津贴。这是你这个月的工资。你想做什么？你想把时间花在什么上面？在这个未来的世界里你会做什么？

(01:11:57) Karina Nguyen

English:

I've been thinking about this a lot times. I feel like I have a lot of jobs options. I would love to be a writer, I think. I think that would be super cool. You should write short stories, sci-fi stories, novels. I really like art history, so you know those conservationists in the museums who just try to preserve art paintings, but just painting through a long day?

中文翻译:

我经常思考这个问题。我觉得我有很多职业选择。我想我会很想当一名作家。我认为那会超级酷。写短篇小说、科幻故事、长篇小说。我也非常喜欢艺术史，你知道博物馆里那些试图保护艺术画作的文物修复师吗？就是整天都在画画修复的那种？

(01:12:28) Lenny Rachitsky

English:

Mm-hmm.

中文翻译:

嗯。

(01:12:29) Karina Nguyen

English:

I think that would be really cool to do. Yeah.

中文翻译:

我觉得做那个会非常酷。是的。

(01:12:36) Lenny Rachitsky

English:

That sounds beautiful.

中文翻译:

听起来很美好。

(01:12:36) Karina Nguyen

English:

I don't know.

中文翻译:

我也不知道。

(01:12:39) Lenny Rachitsky

English:

What I'm hearing is you need to Nerf these models to not get very good at writing so that you can continue... Although at that point you don't need to do it from... You don't need people to buy it, you're just doing it for fun, so it doesn't even matter if they're incredibly good at writing or art conservation. Oh man, what an episode of our conversation. What a wild time we're living in. Karina, thank you so much for being here. Two final questions. Where can folks find you online if they want to reach out and follow up on anything? And how can listeners be useful to you?

中文翻译:

我听出来的是，你需要削弱（Nerf）这些模型，让它们在写作方面别太出色，这样你就能继续……尽管到那时你不需要为了生计而写，不需要别人买，你只是为了好玩而写，所以即使它们在写作或艺术保护方面极其出色也无所谓。天哪，这真是一次精彩的对话。我们生活在一个多么疯狂的时代。Karina，非常感谢你能来。最后两个问题：如果大家想联系你或跟进任何事情，可以在哪里找到你？听众们能为你提供什么帮助？

(01:13:06) Karina Nguyen

English:

You can find me, I'm on Twitter it's KarinaNguyen. You can also shoot me an email on my website. And my team is hiring and so I'm looking for research engineers, research scientists, as well as machine learning engineers, people who come from product engineers who want to learn model training. I'm actually hiring for my team. My team is called Frontier Product Research, and we train models, we develop new methods but for product oriented outcomes.

中文翻译:

你可以在 Twitter 上找到我，账号是 KarinaNguyen。你也可以通过我的网站给我发邮件。我的团队正在招聘，我正在寻找研究工程师、研究科学家以及机器学习工程师，还有那些想学习模型训练的产品工程师。我实际上正在为我的团队招人。我的团队叫“前沿产品研究”（Frontier Product Research），我们训练模型、开发新方法，但目标是面向产品的结果。

(01:13:38) Lenny Rachitsky

English:

What a place to work. Holy moly. What's the best way for people to apply for these very lucrative roles?

中文翻译:

那真是一个工作的好地方。天哪。人们申请这些高薪职位的最佳方式是什么？

(01:13:46) Karina Nguyen

English:

I think you can shoot me a DM on Twitter.

中文翻译:

我想你可以直接在 Twitter 上给我发私信（DM）。

(01:13:49) Lenny Rachitsky

English:

Okay.

中文翻译:

好的。

(01:13:51) Karina Nguyen

English:

Or I'm yet to create a job description for them.

中文翻译:

或者我还没给这些职位写好职位描述（JD）。

(01:13:51) Lenny Rachitsky

English:

Okay. This is the job description.

中文翻译:

好的。这段对话就是职位描述了。

(01:13:58) Karina Nguyen

English:

Or you can apply into post training team. Yeah.

中文翻译:

或者你可以申请加入后训练（post training）团队。是的。

(01:13:58) Lenny Rachitsky

English:

Okay. You're going to get a flood of DMs. I hope you're prepared. Karina, thank you so much for being here. This was incredible.

中文翻译:

好的。你会收到海量的私信。希望你做好了准备。Karina，非常感谢你能来。这次谈话太棒了。

(01:14:03) Karina Nguyen

English:

Thank you so much, Lenny.

中文翻译:

非常感谢，Lenny。

(01:14:05) Lenny Rachitsky

English:

Bye, everyone.

中文翻译:

大家再见。

(01:14:05) Karina Nguyen

English:

It was fun.

中文翻译:

很有趣。

(01:14:09) Lenny Rachitsky

English:

Thank you so much for listening. If you found this valuable, you can subscribe to the show on Apple Podcasts, Spotify, or your favorite podcast app. Also, please consider giving us a rating or leaving a review as that really helps other listeners find the podcast. You can find all past episodes or learn more about the show at LennysPodcasts.com. See you in the next episode.

中文翻译:

非常感谢您的收听。如果您觉得这期节目有价值，可以在 Apple Podcasts、Spotify 或您喜欢的播客应用中订阅本节目。此外，请考虑给我们评分或留下评论，这能极大地帮助其他听众发现这个播客。您可以在 LennysPodcasts.com 找到所有往期节目或了解更多关于本节目的信息。下期节目见。