HAMEL HUSAIN SHREYA SHANKAR

LENNY'S PODCAST

BILINGUAL TRANSCRIPT

ORIGINAL BY

Lenny Rachitsky

@lennysan • x.com/lennysan

ANALYSIS BY

@Penny777 • x.com/penny777

Generated: 2026-01-16

Hamel Husain & Shreya Shankar - 双语对照

Lenny's Podcast: Hamel Husain & Shreya Shankar - The World of Evals

Key Conversation Segments | 核心对话片段

(00:00:00) Lenny Rachitsky

English:

To build great AI products, you need to be really good at building evals. It's the highest ROI activity you can engage in.

中文翻译:

要打造出色的 AI 产品，你必须非常擅长构建评估系统（Evals）。这是你能参与的投资回报率（ROI）最高的活动。

(00:00:05) Hamel Husain

English:

This process is a lot of fun. Everyone that does this immediately gets addicted to it. When you're building an AI application, you just learn a lot.

中文翻译:

这个过程非常有意思。每个尝试过的人都会立刻上瘾。在构建 AI 应用的过程中，你会通过这个环节学到很多东西。

(00:00:12) Lenny Rachitsky

English:

What's cool about this is you don't need to do this many, many times. For most products, you do this process once and then you build on it.

中文翻译:

这件事最酷的地方在于，你不需要重复做很多次。对于大多数产品来说，你只需要完成一次这个流程，然后在此基础上不断完善即可。

(00:00:18) Shreya Shankar

English:

The goal is not to do evals perfectly, it's to actionably improve your product.

中文翻译:

评估（Evals）的目标不是追求完美，而是为了能切实地改进你的产品。

(00:00:23) Lenny Rachitsky

English:

I did not realize how much controversy and drama there is around evals. There's a lot of people with very strong opinions.

中文翻译:

我之前没意识到围绕“评估”会有这么多争议和戏剧性的冲突。很多人对此持有非常强烈的观点。

(00:00:28) Shreya Shankar

English:

People have been burned by evals in the past. People have done evals badly, so then they didn't trust it anymore, and then they're like, "Oh, I'm anti evals."

中文翻译:

人们过去在评估上吃过亏。因为之前做得不好，导致他们不再信任这套东西，然后就会说：“噢，我是反对评估派。”

(00:00:36) Lenny Rachitsky

English:

What are a couple of the most common misconceptions people have with evals?

中文翻译:

关于评估，人们最常见的几个误区是什么?

(00:00:39) Hamel Husain

English:

The top one is, "We live in the age of AI. Can't the AI just eval it?" But it doesn't work.

中文翻译:

排在第一位的是：“我们都处在 AI 时代了，难道不能让 AI 直接来评估吗？”但事实证明，这行不通。

(00:00:45) Lenny Rachitsky

English:

A term that you used in your posts that I love is this idea of a benevolent dictator.

中文翻译:

你在文章中用过一个我非常喜欢的词，就是“仁慈的独裁者”（Benevolent Dictator）这个概念。

(00:00:49) Hamel Husain

English:

When you're doing this open coding, a lot of teams get bogged down in having a committee do this. For a lot of situations, that's wholly unnecessary. You don't want to make this process so expensive that you can't do it. You can appoint one person whose taste that you trust. It should be the person with domain expertise. Oftentimes, it is the product manager.

中文翻译:

在进行这种开放式编码（Open Coding）时，很多团队会陷入“委员会决策”的泥潭。在很多情况下，这完全没有必要。你不想让这个过程变得昂贵到无法执行。你可以指定一个你信任其品味的人，这个人应该具备领域专业知识。通常情况下，这个人就是产品经理（PM）。

(00:01:09) Lenny Rachitsky

English:

Today, my guests are Hamel Husain and Shreya Shankar. One of the most trending topics on this podcast over the past year has been the rise of evals. Both the chief product officers of Anthropic and OpenAI shared that evals are becoming the most important new skill for product builders. And since then, this has been a recurring theme across many of the top AI builders I've had on. Two years ago, I had never heard the term evals. Now it's coming up constantly. When was the last time that a new skill emerged that product builders had to get good at to be successful?

中文翻译:

今天的嘉宾是 Hamel Husain 和 Shreya Shankar。过去一年里，本播客最热门的话题之一就是“评估”（Evals）的兴起。Anthropic 和 OpenAI 的首席产品官都曾分享过，评估正成为产品构建者最重要的核心新技能。从那时起，这成了我采访过的许多顶尖 AI 构建者反复提及的主题。两年前，我甚至没听过“Evals”这个词，而现在它无处不在。上一次出现这种“产品经理必须掌握才能成功”的新技能是什么时候?

(00:05:07) Lenny Rachitsky

English:

I want to use this conversation to basically help people understand this space deeply, but let's start with the basics. Just what the heck are evals? For folks that have no idea what we're talking about, give us just a quick understanding of what an eval is, and let's start with Hamel.

中文翻译:

我想通过这次对话帮助大家深入理解这个领域，但我们先从基础开始。到底什么是评估（Evals）？对于那些完全不知道我们在说什么的人，请简要解释一下什么是评估，我们先从 Hamel 开始。

(00:05:49) Hamel Husain

English:

Sure. Evals is a way to systematically measure and improve an AI application, and it really doesn't have to be scary or unapproachable at all. It really is, at its core, data analytics on your LLM application and a systematic way of looking at that data, and where necessary, creating metrics around things so you can measure what's happening, and then so you can iterate and do experiments and improve.

中文翻译:

好的。评估（Evals）是一种系统地衡量和改进 AI 应用的方法，它其实一点也不可怕，也不是高不可攀。它的核心本质是对你的大语言模型（LLM）应用进行数据分析，用系统化的方式观察数据，并在必要时围绕各项指标建立度量标准，以便衡量现状，进而进行迭代、实验和改进。

(00:08:04) Lenny Rachitsky

English:

So the idea of evals, essentially, is to build a set of tests that tell you, how often is this agent doing something wrong that you don't want it to do? And there's a bunch of ways you could define wrong. It could be just making up stuff. It could be just answering in a really strange way. The way I think about evals, and tell me if this is wrong, just simply is like unit tests for code.

中文翻译:

所以评估的核心理念，本质上是构建一套测试，告诉你：这个智能体（Agent）做出你不希望看到的错误行为的频率是多少？“错误”有很多种定义方式，可能是胡编乱造（幻觉），也可能是回答方式很奇怪。我把评估理解为代码的“单元测试”（Unit Tests），告诉我这是否准确？

(00:08:35) Shreya Shankar

English:

I like what you said first, which is we had a very broad definition. Evals is a big spectrum of ways to measure application quality. Now, unit tests are one way of doing this. Maybe there are some non-negotiable functionalities that you want your AI assistant to have, and unit tests are going to be able to check that. Now, maybe you also, because these AI assistants are doing such open-ended tasks, you kind of also want to measure how good are they at very vague or ambiguous things like responding to new types of user requests... So I would say, overall, unit tests are a very small part of that very big puzzle.

中文翻译:

我喜欢你前半部分的说法，我们有一个很宽泛的定义。评估是衡量应用质量的一系列手段。单元测试只是其中一种方式。也许你的 AI 助手有一些不可逾越的功能底线，单元测试可以检查这些。但由于 AI 助手处理的是非常开放的任务，你还需要衡量它们在处理模糊或歧义问题时的表现，比如响应新型用户请求……所以总的来说，单元测试只是这个宏大拼图中的一小部分。

(00:10:06) Hamel Husain

English:

Yeah, let me just set the stage for it a little bit. So to echo what Shreya said, it's really important that we don't think of evals as just tests. There's a common trap that a lot of people fall into because they jump straight to the test like, "Let me write some tests," and usually that's not what you want to do. You should start with some kind of data analysis to ground what you should even test, and that's a little bit different than software engineering where you have a lot more expectations of how the system is going to work. With LLMs, it's a lot more surface area. It's very stochastic (random), so you kind of have a different flavor here.

中文翻译:

是的，我来做个铺垫。呼应 Shreya 所说的，非常重要的一点是：不要把评估仅仅看作是“测试”。很多人会掉进一个陷阱，直接跳到测试环节说：“让我写几个测试用例吧”，但这通常不是你该做的。你应该先从某种数据分析开始，以此来确定你到底应该测试什么。这和传统软件工程不同，在传统工程中你对系统如何运行有明确预期。而大模型（LLM）的接触面更广，且具有很强的随机性（Stochastic），所以这里的处理方式完全不同。

(00:13:29) Hamel Husain

English:

So what you see here on the screen, this is logs from the application... It's called a trace, and it's just the engineering term for logs of a sequence of events. The concept of a trace has been around for a really long time, but it's especially really important when it comes to AI applications. And so we have all the different components and pieces and information that the AI needs to do its job, and we are logged all of it and we're looking at a view of that.

中文翻译:

你在屏幕上看到的是应用的日志……这被称为“追踪”（Trace），是工程界对一系列事件日志的术语。追踪的概念已经存在很久了，但在 AI 应用中它变得尤为重要。我们记录了 AI 完成工作所需的所有不同组件、片段和信息，现在我们正在查看这些信息的视图。