

谷歌真正的“扫地僧”出山了：Gemini灵魂杰夫·迪恩万字访谈，揭示2026年AI的三个残酷真相

哲学园 2026年2月20日 11:40 加拿大

以下文章来源于New基地，作者🐱



New基地

关注基地，拿回大脑的控制权。第一时间同步马斯克、OpenAI、Gemini 等顶尖大佬...

作为谷歌AI的精神图腾，杰夫·迪恩很少如此详尽地剖析未来的技术路线。

“如果我有更多的时间，我会写一封更短的信。”

在Latent Space的最新访谈中，谷歌首席科学家、被称为“代码之神”的杰夫·迪恩（Jeff Dean）引用了帕斯卡尔的这句名言，来形容未来AI编程的终极形态。

在这个长达一小时的硬对话中，杰夫·迪恩像一个看着未来已经发生的工程师，冷静地拆解了2025-2026年的AI路线图。



如果你还在为现在的AI模型感到惊艳，或者为自己的编程饭碗感到焦虑，看看杰夫·迪恩抛出的这些“暴论”，你可能会感到一种窒息的紧迫感。

1. 关于速度与思考

“每秒生成1万个token有意义吗？绝对有。因为我们需要思维链推理。在给你那1000行最终代码之前，模型可能需要在后台进行9000个token的深度思考和自我验证。”

2. 关于智能的贬值

“我们已经能让下一代的Flash模型（轻量版），达到甚至超过上一代Pro模型（旗舰版）的水平。这是一个我们会持续遵循的趋势。”

3. 关于“无限上下文”的真相

“真正的长上下文不是100万或200万token，而是我们要制造一种幻觉：让你觉得模型正在同时阅读万亿级别的token，正在阅读整个互联网。”

4. 关于提示词工程的本质

“好的提示词工程，本质上就是极其高明的‘高管指令’（Executive Communication）。你要像写内部备忘录一样字斟句酌，因为你在指挥一个超级大脑。”

01. 智能的“通货膨胀”比你想象得更快



我们现在为了使用最强的模型（比如GPT-4或Claude3.5Sonnet），往往要支付高昂的API费用或者是忍受缓慢的推理速度。

但杰夫·迪恩提出了一个极其残忍的“技术下放定律”：

今天的顶配，就是明年的地摊货。

谷歌的策略非常明确：利用“蒸馏”（Distillation）技术，把那个庞大、昂贵、缓慢的“超大杯”模型的能力，压缩进那个极其便宜、极速的Flash模型里。杰夫·迪恩明确表示，每一代新的Flash模型都会追平甚至超越上一代的Pro模型。

这意味着什么？这意味着如果你现在的商业模式是建立在“因为我用了昂贵的大模型，所以我比你强”的基础上，那你很快就会死掉。因为你的竞争对手明年可以用十分之一的成本，用Flash模型做到同样的事。

未来的竞争，不是看谁的模型更强，而是看谁能把“最强智能”白菜价化。

02. 速度本身就是智能



为什么杰夫·迪恩如此执着于 TPU 和 每秒1万token 的推理速度？

很多人以为快只是为了用户体验好，不卡顿。错。快是为了让AI“想”得更久。

现在的AI模型（如DeepSeekR1或OpenAI o1），之所以强大，是因为它们在回答你之前，会进行漫长的“思维链”推理（Chain of Thought）。

试想一下，如果模型每秒只能蹦出100个字，让它进行复杂的逻辑推演，用户得等上几分钟，没人受得了。但如果每秒能蹦出1万个字呢？

模型可以在一眨眼的时间内，在后台把这个问题拆解、反思、自我纠错10次，产生9000个token的“思考过程”，然后只把那个最完美的1000个token的答案吐给你。

当速度快到一定程度，量变就会引起质变。极致的低延迟（Latency），就是通往通用人工智能（AGI）的物理钥匙。

03. 程序员的“死亡”与“高管化”



这可能是整个访谈中对开发者冲击最大的一点。

杰夫·迪恩谈到了未来编程的形态。他不再谈论IDE、不再谈论具体的语法，他谈论的是“50个实习生”。

他打了一个比方：未来的软件开发，就像是你一个人带着50个极其聪明、但需要明确指令的实习生（AI Agent）。

· 以前的编程：你是泥瓦匠，你需要亲自把每一块砖（代码）砌上去。

· 未来的编程：你是项目经理，你需要写的是“规格说明书”（Spec）。

在这个世界里，写代码的能力不再稀缺，稀缺的是“清晰描述需求”的能力。

正如杰夫·迪恩所说，传统的软件工程教育强调写文档、写Spec，但没人爱写。但在AI时代，你的Spec写得有多烂，AI产出的代码就有多烂。

提示词工程（Prompt Engineering）这个词太低级了。未来，这叫“高管沟通术”。你的一句话，需要调度50个Agent去协作，如果你逻辑不清，整个系统就会崩塌。

04. 你要么做高管，要么被淘汰



访谈的最后，杰夫·迪恩描绘了一个“个人化全知模型”的愿景。

未来的AI，不是一个冷冰冰的聊天框，而是一个拥有你所有数据的第二大脑。它读过你所有的邮件、看过你所有的照片、知道你所有的代码库。

在这个新世界里，技术壁垒正在以光速被夷平。硬件在变得更强，模型在变得更便宜。

对于每一个科技从业者来说，警钟已经敲响：别再沉迷于“手搓代码”的快感了。学会像杰夫·迪恩说的那样，去思考架构，去打磨需求，去指挥那支即将到来的、由硅基生物组成的“千军万马”。

因为在2026年，要么你成为那个下达指令的“高管”，要么，你就是那个被Flash模型取代的“昂贵劳动力”。

附：杰夫·迪恩（Jeff Dean）访谈核心问答整理



为了方便大家深入理解，我将本次Latent Space访谈中JeffDean的精彩问答进行了精华整理：

○ Q1：谷歌如何在“探索技术前沿”和“实际部署”之间做平衡？

JeffDean：我们两手都要抓。我们需要前沿模型（Frontier Model）来探索可能性的边界，比如解决复杂的数学难题。但同时，我们需要通过“蒸馏”（Distillation）技术，把这些前沿能力下放到更小、更快、更便宜的模型（Flash）中。没有前沿模型，就无法蒸馏出强大的小模型。

○ **Q2: 你提到的“蒸馏”到底有多重要?**

JeffDean: 非常重要。它的核心优势在于,你可以利用大模型产生的“软标签”(Logits),而不仅仅是最终答案,来训练小模型。这让我们连续几代都做到了:下一代Flash模型的性能,追平甚至超越了上一代的Pro模型。

○ **Q3: 现在的AI基准测试 (Benchmarks) 还有意义吗?**

JeffDean: 公开的基准测试一旦被刷到95%以上的准确率,意义就很难说了,而且还有数据泄露的问题。我们内部更看重那些目前只能拿到10-30%分数的测试集,这才是改进的方向。一旦分数太高,我们就该寻找下一个更难的挑战了。

○ **Q4: Gemini的200万token上下文已经很长了,未来会更长吗?**

JeffDean: 现在的长上下文(1M-2M)很有用,但还不够。真正的终局是让模型给人一种“它可以关注整个互联网”或“它可以关注你一生所有数据”的幻觉。这不可能靠单纯扩大上下文窗口(那是平方级增长)来实现,需要新的检索和层级化系统,从万亿token中瞬间找到你需要的那100个。

○ **Q5: 关于多模态,有什么是我们忽视的?**

JeffDean: 除了听、说、看,AI还应该具备“非人类”的感官。比如激光雷达(LIDAR)、核磁共振(MRI)数据、基因组数据。让模型接触这些模态,即使只是一点点,也能极大拓展它对世界的理解。

○ **Q6: 在AI时代,编程会变成什么样?**

JeffDean: 这就像你管理50个实习生。你不再自己写每一行代码,而是要学会写非常清晰的“规格说明书”(Spec)。以前程序员都不爱写文档,但现在,文档写得好不好直接决定了AI产出的代码质量。这实际上把“提示词工程”提升到了“高管沟通”(Executive Communication)的层次。

○ **Q7: 你对未来有什么预测?**

JeffDean:

1. 极度个性化的模型: 模型将获得权限访问你的邮件、文档、照片等所有数据,成为真正懂你的第二大脑。
2. 超低延迟的深度推理: 随着硬件进步,我们可能实现10,000token/秒的推理速度。这意味着模型可以在几秒钟内进行海量的“思维链”推导,生成9000个思考token,只为了给你一个完美的答案。用帕斯卡尔的话说:“如果我有更多时间,我会写一封更短的信。”AI有了速度,就有了思考的“时间”。

