

Why OpenAI and Anthropic Won't Win the App Layer | Glean Founder – 图文解说

一张关键画面对应一段完整解说，按原视频时间推进。

01. 从模型优势谈到创业者的驱动力 (00:03-01:34)



场景 1

完整内容

现在，90% 甚至更高比例的使用场景，都已经可以被很多很多不同的模型完整处理，其中也包括开源模型。Glean 的创始人是过去十年里最杰出的技术人物之一。他此前创办了 Rubrik，而 Rubrik 显然已经非常成功地上市，也成为了一家很出色的上市公司。现在他又创办了 Glean，这也是一家非常优秀的公司，已经拿到了 Kleiner Perkins 以及许多其他顶级投资人的投资。对于几乎所有不做前沿模型训练的 AI 公司来说，模型公司都应该被看成是一项巨大的资产。一旦你进入消费和使用层，那里并不存在天然的捆绑优势。你必须多做 10 倍的工作，才能从客户那里拿到同样规模的收入。准备开始吧。Arvind，我对这次对话非常期待。我们有一位共同的朋友，Mimoun，他对你赞不绝口。我认为他是这个时代最伟大的“连接者”之一，所以我真的很期待这次访谈。谢谢你来参加。

对方回应说，谢谢邀请。主持人接着说，他觉得创业者通常分成两类：一类是因为“赢”的快感而兴奋，是追逐胜利本身；另一类则是因为害怕失败，被这种对失败的恐惧所驱动。他问 Arvind 属于哪一种。Arvind 说，这是个好问题，他觉得自己大概是后者。他总是在担心什么事情会出错，而这也

让他夜里睡不着。主持人接着说，他很喜欢这种“只有偏执狂才能生存下来”的感觉，并追问他是不是一直都是这样。

画面说明

这一帧里，画面给到 Glean 创始人的近景，他穿着印有 glean 字样的浅色 T 恤，身体前倾，对着麦克风认真讲话，字幕正好落在“当你走向消费层”这句话上。这个视觉状态和这段内容是对齐的：前半段由主持人做高密度开场，先把嘉宾放进“连续创业成功者”和“AI 应用层重要判断者”的位置，再抛出一个关于创业心理的问题；后半段则由嘉宾接住，明确表示自己更多是被“担心出错”的心态驱动。画面中的专注神情强化了这层含义，也把讨论从宏观的模型与应用层竞争，顺势引到这位嘉宾本人的经营判断和风险意识上。

02. 从零开始的创业心态与 Glean 的定位 (01:34-03:04)



场景 2

完整内容

主持人接着问：你一直都是这样吗？Arvind 回答说，基本上是的。主持人说，即便考虑到你已经取得了这么多成功，这还是很有意思。比如 Rubrik 就是一次非凡的成功，如今已经是一家上市公司，而你是联合创始人之一。Arvind 回答说，是的。主持人继续追问，这种心态不会随着时间改变吗？Arvind 说，不会。因为第一，每次你做一家新公司，或者启动一个新项目，在他看来，本质上都像是重新开始。你当然会带着过去的一些宝贵经验，但那仍然是一个全新的世界、全新的环境。

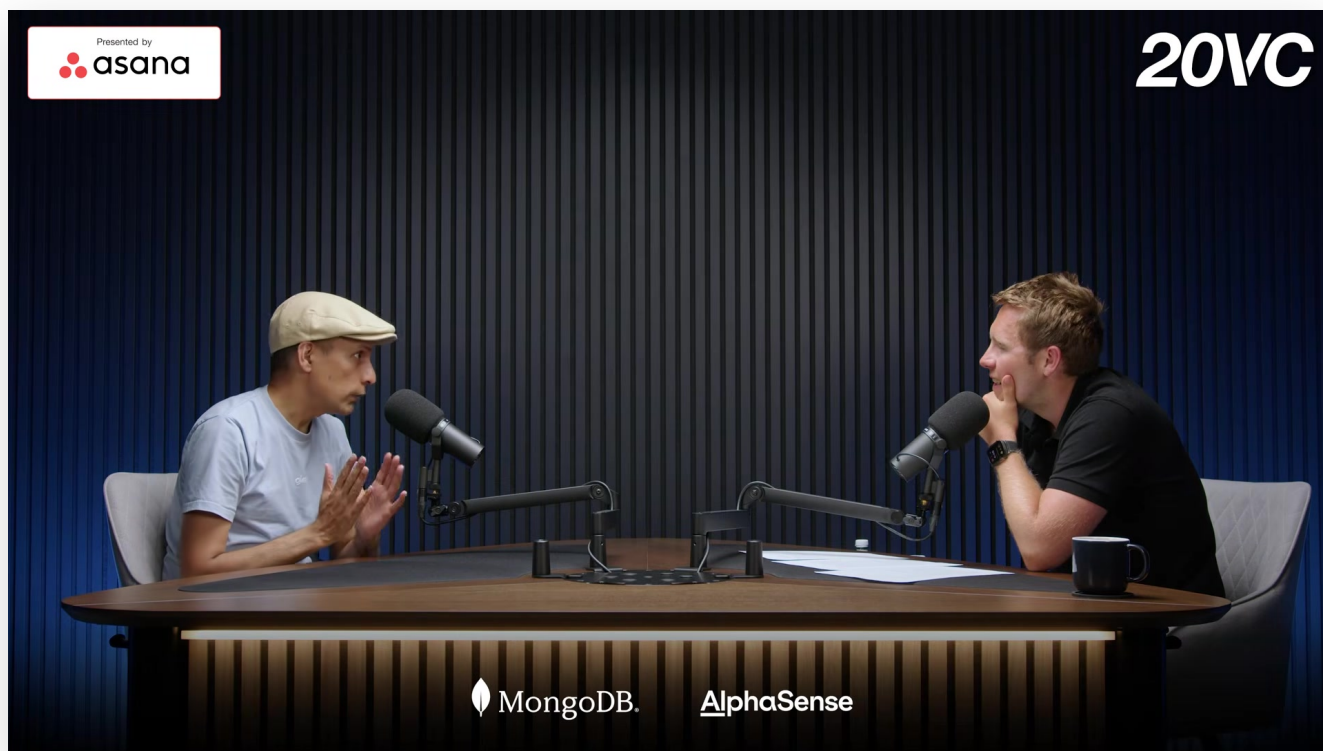
他接着说，想想 Glean 就知道了，它和 Rubrik 在几乎所有方面都根本不同。尤其是在 AI 这个世界里，你必须这样思考，因为几乎每天都会发生一次新的颠覆。如果你把注意力更多放在继续叠加你已经做出来的东西上，带着一种“我们已经赢了，所以继续加码就行”的心态前进，那么在这个新的 AI 世界里，这样做是不够的。主持人随后转换话题，说对于那些不了解 Glean 的人，能不能请他用 60 秒总结一下 Glean 是什么，以及它是怎么运作的。

Arvind 回答说，Glean 是一家企业 AI 公司。公司最初是从企业搜索做起的，也就是帮助员工快速找到他们需要的信息，而这些信息往往散落、埋藏在公司内部 100 个甚至上千个不同系统之中。这是他们起步时的样子，可以把它理解成“工作场景里的 Google”。但随着 AI 模型不断变强，它后来逐步演化成了一个 AI 平台。所以今天如果要理解 Glean，第一种理解方式是：它像是把 ChatGPT、Claude、Gemini 这些能力都结合进一个产品体验里的平台。

画面说明

这一帧仍然是 Glean 创始人的中近景，他一边说话一边抬手比划，桌面和麦克风的播客布景很稳定，左上角有 Presented by asana，右上角有 20VC 标识。这种没有切走的单人镜头，强化了这一段“由嘉宾完整定义自己公司”的结构：先解释为什么连续创业也不能吃老本，再顺势给出 Glean 的 60 秒介绍。画面里他不断用手势强调“重新开始”“每天都在被颠覆”“从搜索演化成 AI 平台”，与内容中的核心论点一致，即在 AI 时代，过往成功不能自动迁移，企业产品必须围绕新的技术能力和新的工作入口重新建构。

03. 企业为何既害怕又怀疑前沿模型公司（03:04–04:35）



场景 3

完整内容

Arvind 继续解释说，Glean 可以被看成是把 ChatGPT、Claude、Gemini 这些能力整合进一个产品体验里的统一平台，它像是员工的一个“同事”，并且连接着公司内部的全部上下文，连接着工作在公司里究竟是如何发生的。主持人随后提到，Palantir 的 Alex K 上周在 CNBC 上说，全球最大的那些企业如今比以往任何时候都更怀疑前沿模型提供商。主持人说，你也服务一些规模极大的企业客户，而且客户质量非常高。你同意他的判断吗？这些企业是不是比以往任何时候都更怀疑前沿模型公司？

Arvind 说，这里有两点。第一，它们确实害怕这些公司。这个“害怕”体现在：就像所有软件公司都会担心“我们未来还有没有生意”“模型会不会把一切都吃掉”一样，企业领导者也在担心，他们的核心知识产权、他们的数据、他们的信息，以及他们学习和做事的方式，会不会最终都变成对这些模型提供商的过度技术依赖。所以这种情绪肯定是存在的。

但他接着说，他理解 Alex K 想表达的另一层意思是：AI 在企业里其实并没有真正奏效，但大家又都害怕把这件事公开说出来，因为在当前的氛围里，承认这件事似乎并不是一件“很酷”的事。主持人立刻打断说，在他们转到“AI 在企业里并没有真正起作用”这个问题之前，他想先停一下，因为那可能是整场访谈里最重要的问题之一，不过那应该单独拿出来讲。于是他先把问题收束成另一个更直接的问题：企业害怕前沿模型提供商会不会把它们午饭吃掉，也就是把它们业务直接吞掉，这种担心到底对不对？

Arvind 回答说，是的，某种程度上是对的，当然这取决于你谈的是哪一类企业。是的，他认为……

画面说明

这一帧切到双人全景，左侧是 Glean 创始人，右侧是主持人，两人隔着桌子和各自的麦克风相对而坐。全景镜头把这一段的“来回追问”关系展示得很清楚：嘉宾先定义 Glean 是连接企业上下文的 AI 同事，主持人随即把讨论拉向更尖锐的市场问题，引用 Palantir 高管的公开判断，追问大型企业是否正在同时“害怕”和“怀疑”前沿模型公司。画面中主持人身体前倾、专注倾听，嘉宾双手收拢解释，正好对应这段论证的结构变化，即从产品介绍过渡到企业客户的真实心理，再进一步逼近“AI 是否有效”和“模型公司会不会吞并应用层价值”这两个核心冲突。

04. 代理 workflow 带来的运营依赖 (04:35-06:06)



场景 4

完整内容

Arvind 继续回答说，如果我们谈的是一种从根本上改变人类工作方式的未来，而且如果我们进一步假设，我们今天所做的大部分工作，未来都会由一个 agent 来完成，而这个 agent 又完全由某一家前沿模型公司提供能力，那么从某种意义上说，你就已经把自己相当大一部分运营活动转移给了这些技术提供商。

他强调，这已经不只是“技术依赖”这么简单了，这实际上是一种真正的运营依赖，因为真正替你运转这些 agent 的，就是那些公司。接着他进一步解释工作是如何随着时间积累隐性知识的：如果你看一个任务是怎么完成的，人在第一次做一件事时，也许会把流程记录下来，比如完成某项工作需要哪 10 个步骤；但随着时间推移，人们会不断优化和微调这个流程，其中有很多东西从来不会被正式写下来。你只是通过反复做这项工作，逐渐积累起各种经验，并在未来实时调用这些经验去完成工作。

他说，所有这些制度性的、机构性的学习，最终都会累积到那个执行工作的 agent 身上。所以，如果你并不能自己掌控这个 agent 的运行，如果你也不拥有它在多年执行过程中积累出来的那些学习成果，那么你实际上就会彻底依赖这些 AI 公司，依赖它们来帮你把工作做完。

画面说明

这一帧回到 Glean 创始人的近景，他皱着眉、双手竖起比划，像是在把一个复杂问题切成清晰的边界来讲。这个视觉动作和他的论点高度一致：他不是在泛泛而谈“模型很强”，而是在严肃地区分“技术依赖”和“运营依赖”。画面里只有他一人发言，说明这段是完整展开的核心论证段落。具体论证是，企业一旦把多数工作交给由前沿模型驱动的 agent，连同流程优化过程中那些没有被文档化、但会不断复利的隐性经验，也会一起沉淀到外部 agent 中；如果企业无法自己掌控 agent 的运行和学习成果，就会在业务层面被深度绑定。

05. 企业转向开源的拐点与成本压力 (06:06-07:37)



场景 5

完整内容

Arvind 继续说，如果企业最终要依赖这些 AI 公司来把工作做完，那么今天摆在企业面前就有一个根本问题：它们到底要如何使用这些 AI 技术，同时又保留控制权，并且保留那些随着 AI 一起不断复利、不断累积的学习成果？这些成果应该属于企业自己。

主持人接着问，他是否看到企业客户正在从前沿模型提供商那里把重心转向开源模型。Arvind 说，这件事现在确实正在发生。他认为现在正处在开源的一个真正拐点上。其中一部分原因是，大家一直都在等待开源模型真正变得更好。企业对开源的渴望已经存在很多很多年了。他说，他们接触的企业里，没有哪一家会满足于说：“好吧，我只靠 OpenAI 或 Anthropic 就能把工作做完，这样我就满意了。”每一家企业都希望确保自己能掌握自己的命运，都希望自己可以使用多种模型。

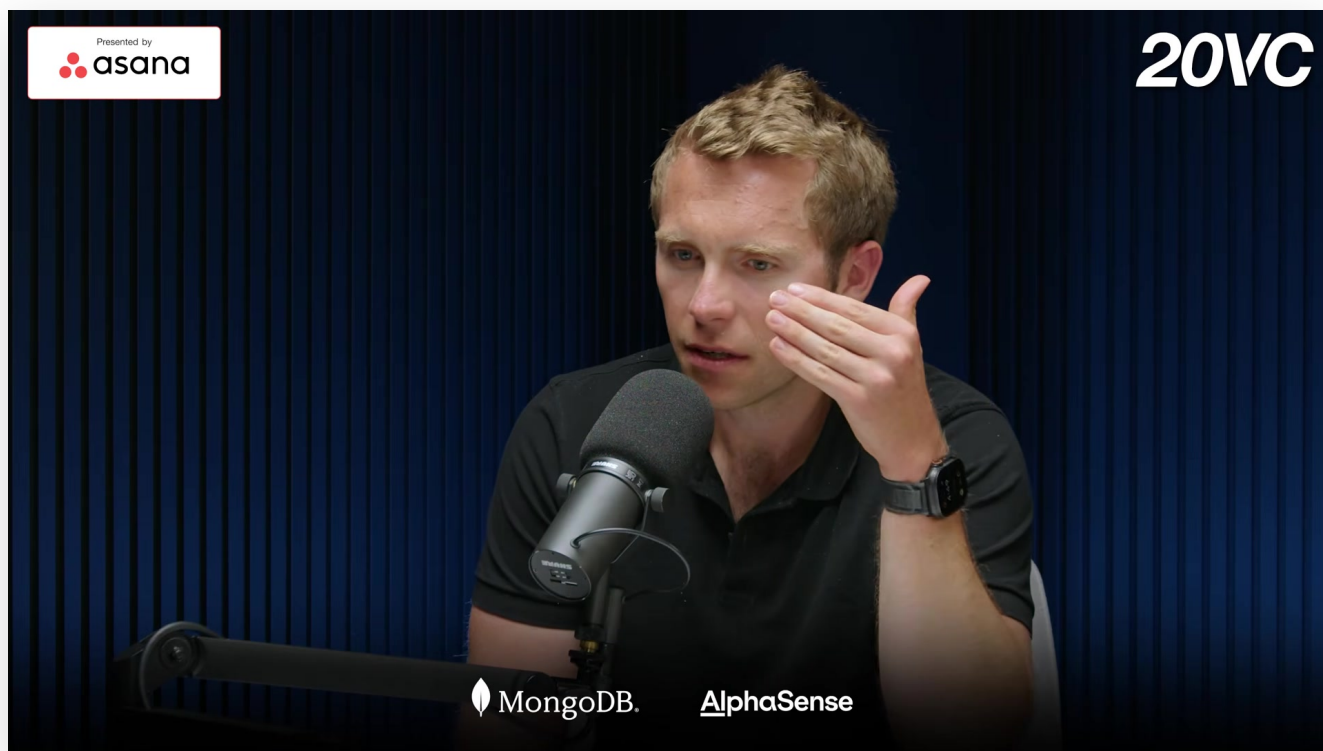
他接着补充说，现在 AI 已经变得非常昂贵，这也是关键原因。你总能听到很多公司这样的故事：它们给 AI 制定了一个年度预算，但结果往往在一两个月之内就把预算跑穿了。主持人插了一句：“CFOs。”Arvind 接着说，对，所以这种情况进一步强化、加速了企业对开源的渴望；再加上现在开

源世界里已经出现了非常好的模型，这两个因素叠加在一起，推动了这次转向。主持人随后追问：企业到底在乎什么？它们在乎的是成本吗？还是所有权？比如数据留在本地部署环境里、模型本身是它们可以看得见、看得清的，这些才是重点吗？到底是哪一个？

画面说明

这一帧仍然是 Glean 创始人在说话，他双手靠近面前，姿态像是在逐层推进一个判断。镜头没有切给主持人，说明这一段重心放在嘉宾对市场变化的解释上。画面与论点的关系很直接：他先把“控制 AI 带来的复利学习成果”设为企业的根本命题，然后回答主持人的追问，明确表示企业现在确实在向开源转向，而且这是一个“真实的拐点”。这里的完整论证包括三层：企业一直想掌握自主权，不愿把命运完全交给 OpenAI 或 Anthropic；AI 成本已经高到年度预算能在一两个月内被迅速打穿；再加上开源模型现在确实已经足够好，所以企业开始更实际地推进多模型和开源路线。

06. 开源驱动力来自成本，垂直包仍然很浅（07:37-09:08）



场景 6

完整内容

Arvind 回答说，就目前来看，推动企业转向开源的主要动力还是成本。从另一个角度看，当然也有一些企业本来就有要求，必须把所有推理负载都保留在自己的私有数据中心之内。AI 刚出现的时候，企业对把数据放到自己控制范围之外这件事要恐惧得多，它们也担心模型公司会用它们的数据继续训练模型。但这种恐惧现在已经没那么强了。如今企业普遍相信，只要自己签的是正确类型的合同，模型公司就会负责任，不会拿企业数据来训练模型。所以，当前的主要驱动力还是成本。

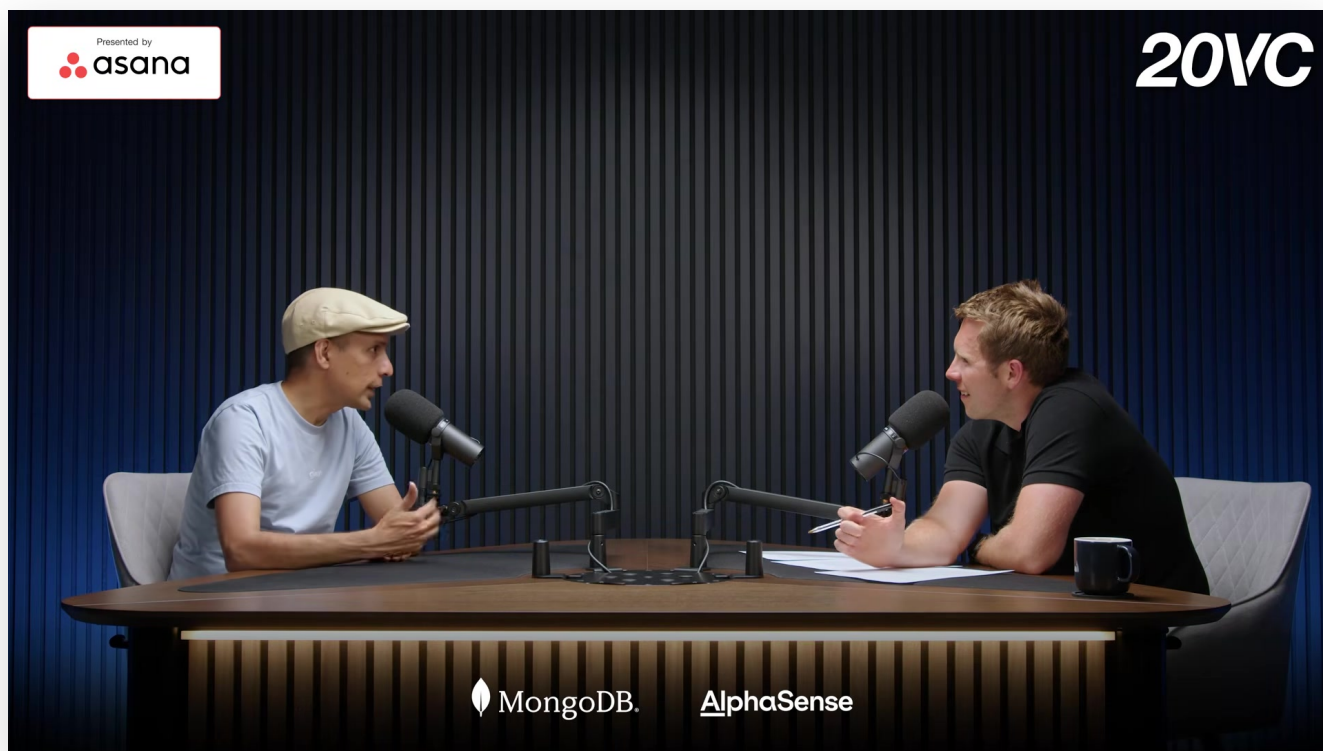
主持人接着说，从你所在的位置看整个竞争版图，每一个跟他聊过的投资人都说他必须问这个明显的问题：你会不会担心 Anthropic 会像他们对 Figma 所做的那样，或者像他们在法律领域做的那样，或者像他们在医疗领域正在做的那样，直接进入你的领域，并蚕食你的业务？Arvind 先回应说，首先，他们应该谨慎一点，先搞清楚 Anthropic 实际上到底对 Figma、法律行业或者金融行业做了什么。它们确实在推出这些垂直领域的套件，但在他看来，这些东西相当浅。

他接着说，据他所知，并没有人因为这些产品，就把自己的工作负载从 Figma，或者更一般地说，从任何其他工具，完整地迁移到 Anthropic 上。那更像是一些净新增的使用场景。

画面说明

这一帧切到主持人的近景，他一只手抬起放在脸侧，身体朝着麦克风前倾，明显是在发出一个更具攻击性的问题。这个画面转换很关键，因为这一段的主导权先短暂回到主持人：他先追问企业究竟在乎成本、所有权还是本地可见性，随后又把问题升级为“Anthropic 会不会直接吃掉你的市场”。虽然画面里是主持人在发问，但这一整段的核心答案来自嘉宾，而且答案分成两个明确部分：第一，企业此刻推动开源的头号原因是成本，而不是最初那种对数据外流和被训练的强烈恐惧；第二，Anthropic 推出的垂直行业包在他看来还比较浅，并没有让用户把工作负载整体从 Figma 或其他现有工具迁过去，因此更像新增尝试，而不是对成熟应用层产品的全面替代。

07. AI 在扩市场，不只是在替代老工具（09:08–10:41）



场景 7

完整内容

他说，很多情况下这其实都是一种净新增的使用，而不是简单把原有工具完全替掉。更准确地说，他认为 AI 正在扩大市场。比如在设计领域，设计师当然仍然会使用 Figma，但那些本来不是设计师的人，现在也开始用 Claude 去做“云设计”这类工作。也就是说，AI 正在把很多事情变得更简单；当一个人并不是某个专业工具的核心专家用户时，他现在也可以借助 Claude，开始完成其中一部分原本只有专业人士才会做的工作。

主持人接着问，所以你不会担心这些模型公司都在强调进入企业、去占据“上下文”这一层吗？

他回答说，不管他们是不是正在这么做，他们事实上都已经在这条路上了，而且 Glean 每天都在面对这种竞争。企业客户经常会问他们：Claude 也可以通过 MCP 连接企业系统，那到底有什么不同？

Glean 能做到什么，而 Claude 做不到什么？所以他们必须反复解释，真正的“上下文”到底是什么，以及为什么构建这种上下文其实非常复杂。换句话说，他们确实在和这些模型公司竞争。甚至他会说，

这些公司可能比其他对手更早开始和他们竞争。因为如果你把 Claude 的 co-work 或 Claude desktop 看成一种应用，那么它最主要的用例一直都是问答。事实上，今天全世界 AI 最大的应用场景，就是信息获取和问答。

最后主持人转向下一个问题：你觉得率先进入市场到底有多重要？他说，这实际上是非常有利的，但也只是……

画面说明

这一帧是双人对谈的宽景，左侧的受访者正边说边用手势解释，右侧主持人前倾倾听，画面里没有任何产品演示或图表，强调的是一场围绕商业逻辑的直接交锋。画面对应的核心论点是：AI 并不只是替代既有软件，而是在把原本不会使用专业工具的人拉进市场；但与此同时，模型公司也已经顺着企业连接、MCP 和问答入口切入应用层。受访者在镜头中主导发言，正是在说明 Glean 为什么既把这种变化视为市场扩张，也承认自己已经与 Claude 这类产品在企业问答和信息获取场景里正面竞争。

08. 先发优势有价值，但绝不是救命符（10:41–12:11）



场景 8

完整内容

他接着前一句说，先发优势确实有帮助，但它不是那种能够永远把你一路带下去的东西。对他们来说，Glean 的确因为“世界上第一家企业 AI 公司”这层身份获得了很多信用。他们是最早把 RAG 带进企业的人，也是最早构建概念式语义搜索的人。因此，这给了他们品牌，也给了他们在这个市场里参赛的资格。即便现在和已经变得非常庞大的 OpenAI、Anthropic 这些巨头相比，他们的体量要小得多，这种先发身份依然是一个巨大的资产。但与此同时，它既不是进入这个市场的必需条件，也不是一个能拯救一切的东西。

主持人随后问：你会怎么建议那些因为担心前沿模型提供商进入自己赛道、晚上睡不着觉的创始人？

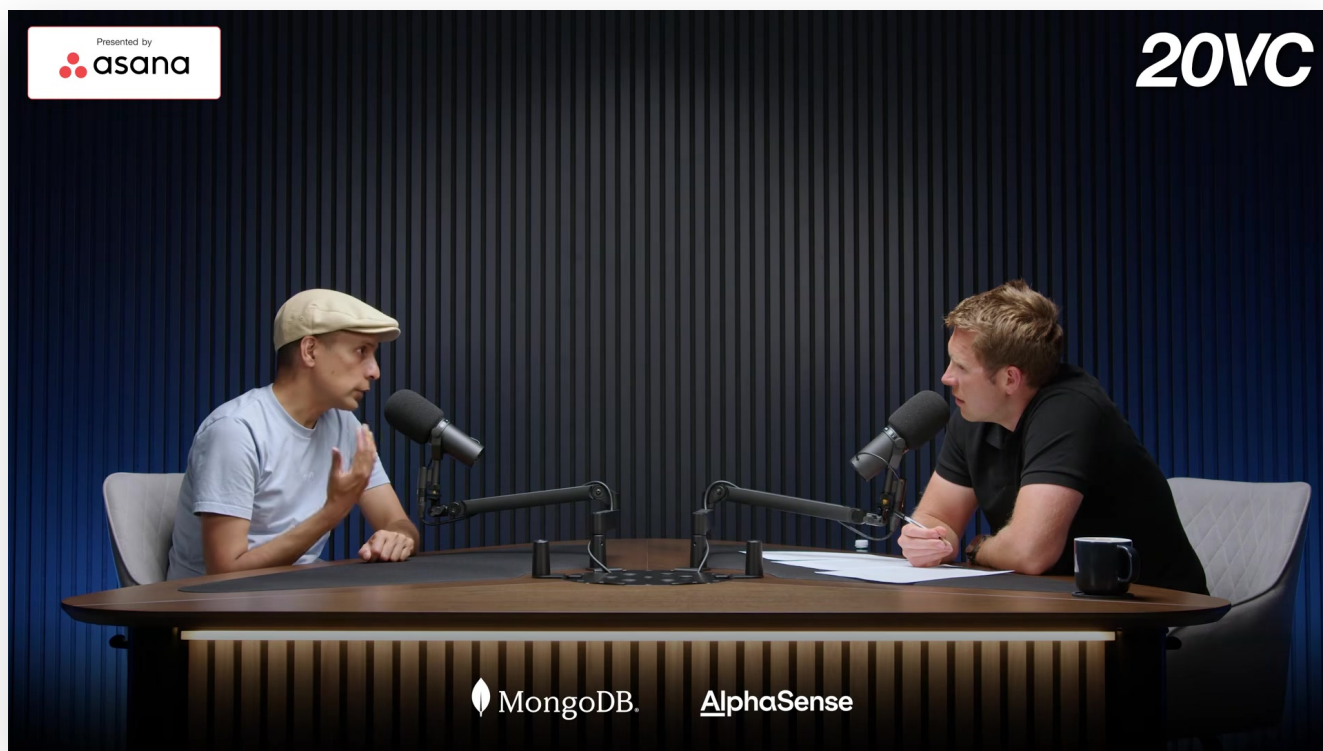
他说，他的建议是完全不要为这件事担心。作为创始人，你真正该做的是解决问题，而不是先去担心。没错，你当然需要始终预判这些模型公司接下来会做什么，也要了解它们当前的能力边界；但是在他看来，对于几乎所有不做前沿模型训练的 AI 公司来说，模型公司更应该被视为一种巨大的资

产，而不是竞争对手。无论是 Anthropic 在做的事、OpenAI 和 Google 在做的事，还是开源社区里正在发生的各种创新，对他们来说其实都是好消息。他们并不会因此焦虑，也不把这些事情首先看成竞争。恰恰相反，这些进展让他们能够……

画面说明

这一帧切到受访者的近景，他闭着眼微微皱眉、双手在胸前比划，像是在认真拆解一个容易被神话的问题。画面强化了这一段的语气：他不是在鼓吹“先发就赢”，而是在给创始人降噪，指出先发优势只是有利条件，不是护城河本身，也不是拯救公司的万能药。镜头里只有他一人发言，因此视觉上也贴合这段内容的主旨，即他以创始人和经营者的角度，反过来主张其他 AI 公司应把 OpenAI、Anthropic、Google 和开源创新首先视为能力供给与市场放大的来源，而不是让自己失眠的威胁。

09. 企业场景下，模型层正在快速商品化（12:11-13:43）



场景 9

完整内容

他把上一句补完：这些模型公司的进展，让他们能够交付出如果没有这份帮助就根本做不出来的产品。

主持人随即追问：当你看到 Anthropic、OpenAI，以及模型层不断上升，再加上新模型推出的速度这么快，尤其是来自中国开源提供商的捆绑式推进时，难道我们不是正在见证模型层的终极商品化吗？

他回答说，至少有一件事已经很清楚，不过要把讨论限定在企业使用场景里。对于企业 use case 来说，90% 甚至更多的场景，现在都已经可以被很多很多不同的模型完整处理，其中也包括开源模型。所以从这个角度看，商品化是明确存在的。事实上，在 Glean，这甚至已经成了他们给客户提供的核心价值之一，也就是成本控制。他们会直接告诉客户：随着用户在他们的平台上完成各种任务，Glean 会替客户选择合适的模型；如果客户接受使用开源模型，那么在他们判断合适、并且能够生成高质量答案的时候，他们就会用开源模型。

主持人接着问：有多少客户是不接受开源模型的？

他说，这件事其实还非常新。因为真正意义上让开源模型在能力上逼近前沿模型，达到只差 3 个月左右的水平，这件事几乎就是最近才发生的，literally 也就一个月前，甚至不到一个月前。他还举例说，GLM 5.2 的出现，是第一次让他们自己的团队都开始觉得，可以放心地把大多数工作负载跑在这个模型上。

画面说明

这一帧回到双人宽景，受访者在左侧抬手强调观点，主持人低头拿笔、身体前倾，明显进入了具体追问的节奏。画面与论点的关系很直接：讨论已经从“该不该怕模型公司”推进到“模型层本身是否正在被商品化”。受访者在这一段给出最硬的判断和数字，说在企业场景中 90% 以上的用例已可由多种模型完成，因此 Glean 的价值开始转向模型编排与成本控制，而不是押注某一个唯一模型；主持人的追问则把话题进一步压到客户是否接受开源，以及这种变化到底新到什么程度。

10. 真正的分歧，是敢不敢用中国模型（13:43–15:14）



场景 10

完整内容

他说，对他们自己的团队来说，这是第一次他们真的觉得，可以安心地把大多数工作负载运行在那个模型上。所以，他们现在还在等待市场最终会怎么反馈。他还补充说，从“开源”本身以及“使用模型”这件事的角度来看，他并不认为大家会有太大问题。真正的问题会是：企业到底能不能接受中国模型。这才是这里唯一真正的问题，而不是开源对闭源的问题。

他进一步追问：既然企业可以掌握所有权，也可以把模型部署在本地 on-prem，又不会把任何东西回传给中国，那为什么不能接受中国模型？

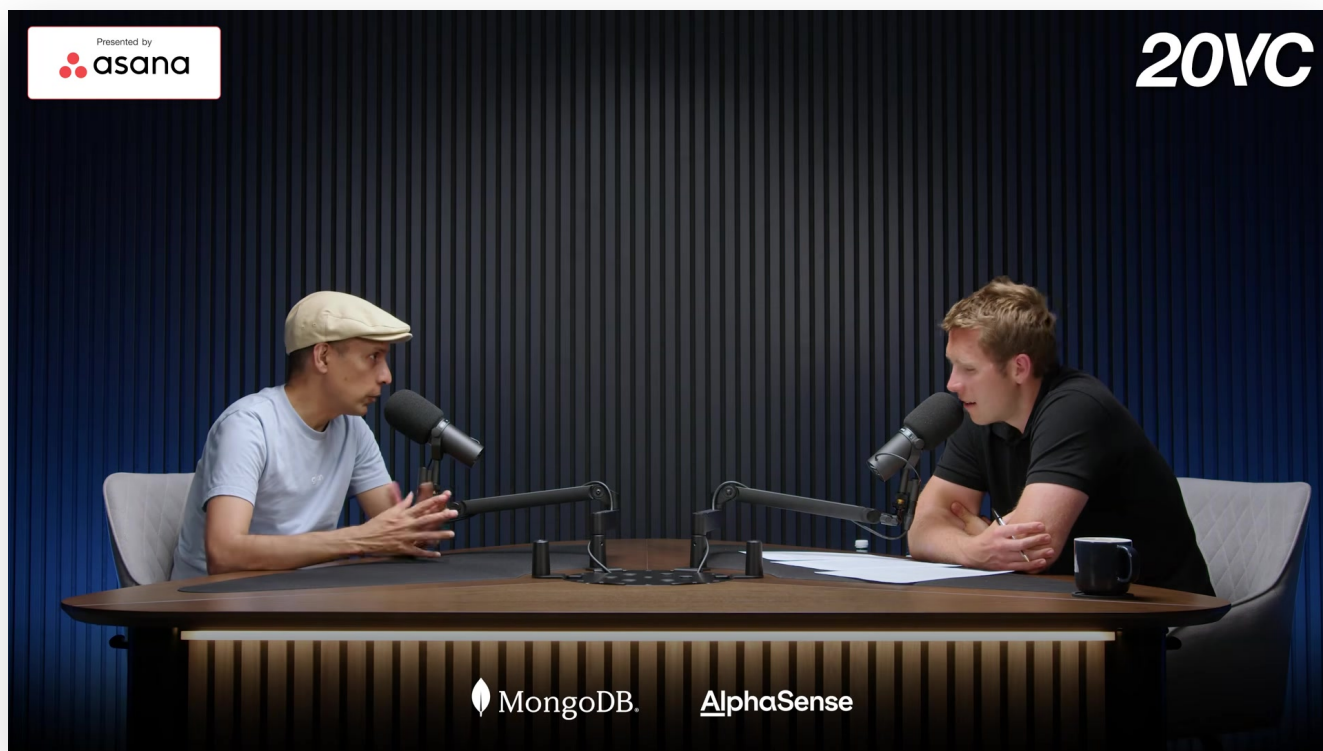
主持人回应说，他觉得这更多还是一种心理舒适度的问题。大家会担心，万一出了什么事怎么办？总会有某种偏执和恐惧，比如会不会存在后门，会不会有某种他们根本无法理解的“魔法式后门”。所以确实会有一些顾虑。除此之外，如果你使用这些模型，而且这件事变得众所周知，那么竞争对手也可能在某些情况下拿这件事来攻击你，诸如此类。所以这里面有很多因素。

不过主持人最后也承认，归根到底，这还是回到一个问题：谁愿意更大胆一些。因为这毕竟是新事物。大型企业终究必须迈出这一步，而最早行动的人会先迈出去，之后它就会逐渐变成一件更正常的事情。主持人也顺带说，他一直做这个节目，也是想搞清楚：如果 90% 的企业工作流都可以用开源模型完成，那么我们是否从一开始就把很多东西定价定错了？

画面说明

这一帧是受访者的单人近景，他一只手抬起、手指张开，像是在把问题拆成几个清晰的层次。视觉上没有任何外部素材干扰，因此重点完全落在他说出的判断上：争论的核心已经不再是“开源还是闭源”，而是企业是否愿意信任中国模型，即使这些模型可以本地部署、数据不回传。随后主持人从风险感知、后门焦虑、竞争对手攻击等角度提出反驳，形成这一段最明显的来回。画面里受访者持续主导发言，也正对应这段论证的主轴，即技术与成本层面已经越来越可行，真正卡住 adoption 的往往是组织心理和政治风险。

11. 模型生意未必像外界想的那样暴利（15:14–16:46）



场景 11

完整内容

主持人问：我们是不是把前沿模型这整个市场定价得完全错了？现在已经是一个非常不同的时代了。

他回答说，他确实觉得，单看模型业务本身，无论先暂时不谈开源，实验室之间本来就已经有大量竞争，而且还会有越来越多公司进入这个领域。所以，在这种激烈竞争之下，即便只是一个三方竞赛，你也会看到相当明显的价格压力。再加上开源的存在，开源模型的价格又便宜一个数量级。因此他甚至听到一些传闻，说 OpenAI 可能会大幅下调模型价格，以回应这些新发展、市场竞争以及开源的压力。所以总体上，他认为单独的模型业务本身，可能并不像大家以为的那么赚钱、那么丰厚。

但他也补充，这些公司现在已经拥有更多东西了。它们不再只是模型公司，不再只有模型业务。

主持人表示，他完全理解这一点，但如果这些公司在那些相邻领域里做的只是一些比较浅的事情，那它们也不太可能因此创造出一万亿美元的收入。接着主持人提到 Dario 的说法。

他回应说，首先，这两家实验室本质上就是非常不同的业务。OpenAI 当然拥有非常出色的消费者产品。而 Anthropic 这边，正在发生的一件有趣的事是……

画面说明

这一帧再次回到双人宽景，但此时主持人低头看着桌上的纸、像是在引用或抛出更尖锐的判断，受访者则坐在对面继续回应。这个构图很适合这一段内容：主持人把问题升级为“整个前沿模型市场是不是被高估、被错价了”，而受访者用竞争、价格压力、开源便宜一个数量级、以及 OpenAI 可能降价的传闻来支撑自己的回答。画面中的两人都在场，也反映出这不是单向陈述，而是关于模型业务盈利性、以及 OpenAI 和 Anthropic 是否已经超出“纯模型公司”范畴的一次来回拆解。

12. Anthropic 在长成平台，微软则靠捆绑施压（16:46–18:16）



场景 12

完整内容

他接着说，正在发生的一件有趣的事情是，人们其实正在构建在这些公司的平台之上。比如说现在看 Anthropic，就有很多人在它的平台上开发自动化、skills，大家还在创建这些 MCP servers，把它们连到自己的内部系统里，再把这一切接到 Claude 上。所以，一个生态系统实际上正在形成。正因为如此，你应该把 Anthropic 看成一家应用层公司，而不只是模型公司。

主持人接着问：如果让你猜 3 年后，你们的工作流里会有多少百分比跑在开源之上？

他说，他们一直在告诉客户，他相信 3 年之内，大多数企业工作负载肯定都会跑在开源模型上，这是非常明确的判断。

随后他又提出另一个模型提供商面临的竞争因素：其实是微软。微软非常擅长做这样一种生意，即产品可能只有别人 70% 那么好，但它把这个产品打包进企业采购 bundle 里，再贴上一个漂亮的标价卖给企业，而且这门生意做得极其成功。

主持人立刻问：那你怎么看来自 Microsoft Copilot 这种捆绑压力的竞争威胁？

他回答说，对他们而言，微软确实是他们最重要的竞争对手之一，而且这种 bundling 策略确实有效。你必须正面去和它对抗。不过幸运的是，市场一直都给 best-of-breed、也就是同类最佳的软件留有空间，而他们的客户正是这样看待 Glean 的。

画面说明

这一帧是受访者的近景，他身体前倾、双手交握，像是在给出一个更完整、更成体系的行业判断。画面与论点的关系非常紧：这一段不再只是谈单个模型的能力或价格，而是谈平台生态、未来工作负载迁移，以及真正的企业分发对手。受访者先解释为什么 Anthropic 已经在通过 automations、skills、MCP servers 和内部系统连接长成应用平台，然后又把时间线拉到 3 年后，明确判断大多数企业工作负载会跑在开源模型上。最后他把竞争焦点落到微软身上，指出真正有效的威胁往往不是最强模型，而是能把“70% 好”的产品通过企业捆绑销售出去的分发机器。

13. 按量计费会削弱捆绑优势吗（18:16–19:47）



场景 13

完整内容

嘉宾说，客户对他们的认知正是如此：如果一家企业想真正获得优秀的搜索产品，或者想搭建一个横向、全面的 AI 平台，他们知道 Glean 在这件事上做得更好。所以，即便企业已经在使用微软打包的整套产品，很多公司仍然愿意在这之上继续投入。

他接着说，另一个会让“捆绑销售”这套策略不再像过去那样有效的因素，是 AI 正在走向按使用量计费的模式。一旦进入按量消费，就不存在天然的捆绑优势了。因为作为一家企业，我完全可以同时采购六个工具，让员工自己选择去哪里完成工作；无论他们最终在哪里工作，我都只需要为那一个具体工作单元付费。所以从长期看，按消费计费这种模式可能会打破捆绑策略。

主持人礼貌地表示自己不太认同，至少在企业市场未必如此。他说，企业会把你纳入一个合规的企业级采购包里。对于世界上最大的那些公司，无论是 VW、Ford、P&G，还是 Tyson Chicken 这种级别的企业，内部都会有审批流程、签核流程，以及一整套合规程序。他还打趣说自己总爱举这些随机公

司当例子。问题在于，这些企业也许愿意把微软批准为一个供应商，但如果突然要去批准 15 个供应商，不只是价格和交易变复杂了，更会带来以前没有的供应商管理问题。

画面说明

这一帧是嘉宾的近景，他戴着浅色帽子、穿着印有 glean 标识的浅蓝 T 恤，正朝着话筒方向认真讲话，画面没有切到主持人，说明此刻主要是他在完整展开论点。这个镜头和内容是匹配的：他在解释为什么即使微软有套装优势，客户仍会把 Glean 视为更强的横向搜索与 AI 平台，并进一步提出“AI 正走向按量消费，所以捆绑会被削弱”的核心判断；而这段论证随后立刻被主持人从企业合规、审批和供应商管理的角度挑战，形成了这一时间段最关键的分歧。

14. 微软免费捆绑比前沿模型更难打 (19:47-21:17)



场景 14

完整内容

主持人刚才的反驳说完后，嘉宾先回应说，这一点确实成立。但他补充，如果你去和那些曾经在微软攻势下吃过亏的公司交流，大多数人都会说，真正的“杀手”其实是定价，因为你很难和“免费”竞争。

随后主持人追问：更凶猛的竞争对手到底是谁，是微软，还是那些前沿模型公司？嘉宾回答说，现在下结论还太早，但微软确实非常强大、非常难缠。就他们自己的销售经验来看，在做客户拓展时，他们更常听到的回答是：“我们已经是微软客户了，也已经拿到 Copilot 了，所以再考虑你们就没什么意义了。”这种说法，比“我已经全面拥抱了某个实验室的产品，所以我不会再做别的选择了”要常见得多。

主持人接着把话题拉回到价值兑现上。他提到前面说过的 Alex 那场访谈，说自己刚才打断了对方，是因为在深入讨论之前，他想先指出一个判断：他认为到了 2026 年下半年和 2027 年，所有人都会停下来问一句，这笔 AI 开支到底有没有产生实际产出，或者有没有带来 ROI。企业究竟从这些投入里

得到了什么投资回报？Alex 的说法会不会是对的，也就是大家最终都会开始追问：“到底怎么回事？我的回报在哪里？”

画面说明

画面仍然是嘉宾的单人近景，他身体微微前倾，双手在胸前收拢，语气像是在谨慎承认主持人的一点看法，同时把话锋重新引回自己的主张。这个镜头下，虽然主持人的提问听得很清楚，但视觉上仍由嘉宾主导，因此更突出这一段的逻辑推进：他承认企业采购流程确实会让多供应商更难，但他坚持认为真正最致命的是微软用已有客户关系和近乎“免费”的 Copilot 捆绑来卡位，而主持人则顺势把讨论推进到 2026 年下半年与 2027 年的 ROI 拷问上。

15. AI 已有局部价值 但交付速度未必更快 (21:17-22:48)



场景 15

完整内容

嘉宾说，如果大家都在问“我的回报在哪里”，他的看法是：今天已经有一些明确实现价值的“口袋地带”。例如客户支持这个垂直场景，就很容易衡量生产率。你完全可以非常具体地说，在你的公司里，一个客服一天原本能处理 10 个 case，现在因为 AI 能处理 12 个。这样一来，生产率提升就能被非常直接地量化出来。

他进一步解释，客服之所以是 AI 表现不错的用例，是因为支持团队花掉的大量时间，本来就是在阅读知识，再把这些知识整理、总结后反馈给客户。从这个角度看，AI 很适合介入。所以，确实存在一些领域，价值实现是清晰的，企业也会因此感到满意。

但他也说，另一些场景就复杂得多。比如现在大部分 AI 开支，其实都花在编程上。如今多数开发者都在用 AI 写代码，已经不是靠手工一行一行写初稿了。从某种意义上你当然可以说，AI 的影响已经很大；但真正的问题是，产品是否因此更快上线了？而在这一点上，他们听到的大多数公司反馈都是

“没有”。也就是说，尽管编码速度显著提升了，产品真正的交付速度并没有同步上升，因为写代码本身只占产品整体交付流程中的一小部分。

画面说明

这张关键帧里，嘉宾依旧是唯一出镜者，皱着眉、靠近麦克风，右手在胸前做动作，像是在把“哪里已经有价值、哪里还很复杂”拆开讲清楚。视觉上没有任何辅助图表，因此这一段的重点完全落在口头论证上：他先用客服从每天 10 个 case 提升到 12 个 case 的例子证明 AI 已经能带来可测量的 ROI，再把镜头中的严肃语气延伸到更难衡量的编程场景，指出“编码更快”并不自动等于“产品交付更快”。

16. 几乎所有初稿代码都由 AI 生成（22:48–24:18）



场景 16

完整内容

主持人追问：那你们自己的产品交付速度提高了吗？嘉宾回答说，这其实很难衡量，这正是难点所在。工程生产率是最难衡量的事情之一，工程岗位本来就是最模糊的一类工作。如果只看某些指标，比如写出来的代码行数，那当然现在比以前多得多；但如果问“我们是不是以更快的节奏在发布功能”，答案是“是，我们确实更快了”，只是这同时也和别的因素有关，比如团队更大了，团队成员的资历和成熟度也比以前更高了，所以很难把这些因素彻底拆开。

尽管如此，他说他们公司当前的选择很明确：继续投资。主持人接着问，现在 Glean 有多少比例的代码是由 AI 写的？嘉宾说，差不多接近 100%。现在几乎没有人再手工写初始代码了，顶多偶尔会有例外。主持人插话打趣说，也许角落里还坐着一个“手工艺人”。嘉宾顺着这个玩笑回应，说是的，总之现在几乎所有代码都是借助 AI 写出来的。

不过他也强调，他们强制执行人工评审。你不能让 AI 一口气生成大量代码，然后不经检查就直接提交进代码仓库。他们不会允许这种做法。

画面说明

画面是嘉宾的近景，他把双手竖在身体前方，像是在给一个模糊问题设边界，这和他此处的说法一致：工程效率并不是一个可以简单用单一指标下结论的问题。镜头里只有他一个人，因此观众会把注意力集中在这组判断上，即“功能发布可能更快了，但原因不只 AI”，以及更醒目的那句“几乎 100% 的初始代码都由 AI 写出”；而他在同一段里又立刻补上一条限制条件，即代码依然必须经过人工评审。

17. 评审成了新瓶颈 但还不能取消 (24:18–25:51)



场景 17

完整内容

嘉宾接着说，他们大概比多数公司更保守。公司内部确实讨论过一个问题：既然 AI 现在能写出这么多代码，真正的瓶颈其实已经从“写代码的人”转移到了“审代码的人”。因此，内部一度有人提议，干脆取消 code review，让代码直接提交进仓库。事实上，他也说现在很多公司确实在这么做。

主持人随即指出，如果你仍然保留严格的代码审查流程，那几乎就削弱了快速代码开发流程的意义。嘉宾回应说，这个判断没错。但他认为，这恰恰说明大家现在仍处在“学习如何正确使用 AI”的阶段，也还在思考它的长期后果。因为如果你用 AI 写代码，确实可以一下子写出一百万行，但接下来要长期维护、理解和管理这些代码，就会变得极其困难。

主持人反问说，可 AI 不就是也能做这些事吗？它也能做重构，也能做安全检查。嘉宾回答说，问题只是它现在还没有好到那个程度。所以，在当前这个阶段，他们宁愿承担代码评审的成本。

画面说明

这一帧里嘉宾一只手贴近胸口，另一只手竖起，像是在做取舍和分层说明，镜头语言很适合承载“速度”和“可维护性”之间的权衡。由于画面仍然只给到嘉宾，视觉焦点完全落在他的解释上：AI 已经把代码生成推到很快，但瓶颈转移到了 review；主持人用“那快还有什么意义”来追问，嘉宾则把论点推进到长期维护风险，强调现在取消人工 review 还为时过早。

18. ROI 的关键在于上下文吞吐 (25:51-27:22)



场景 18

完整内容

嘉宾说，尽管他们仍然保留评审环节，但整体上还是比以前快，因为真正写代码这一部分现在快得多，而且写代码的人自己也会先做第一轮审查。

主持人接着问：当你说 AI 的 ROI 本质上是一个“吞吐量问题”时，这到底是什么意思？嘉宾回答说，首先要确保的一件事，是你能够把正确的上下文提供给这些 AI agent。他解释说，今天大多数企业部署 AI 的方式，往往只是很粗糙地把它直接扔进系统里，再用 MCP servers 之类的方式，把 AI 和企业内部所有系统连接起来。这样一来，无论你想让 AI 完成什么工作，实际上都是在让模型用一种近乎蛮力的方式，自己去摸索、拼装完成任务所需的原始材料。

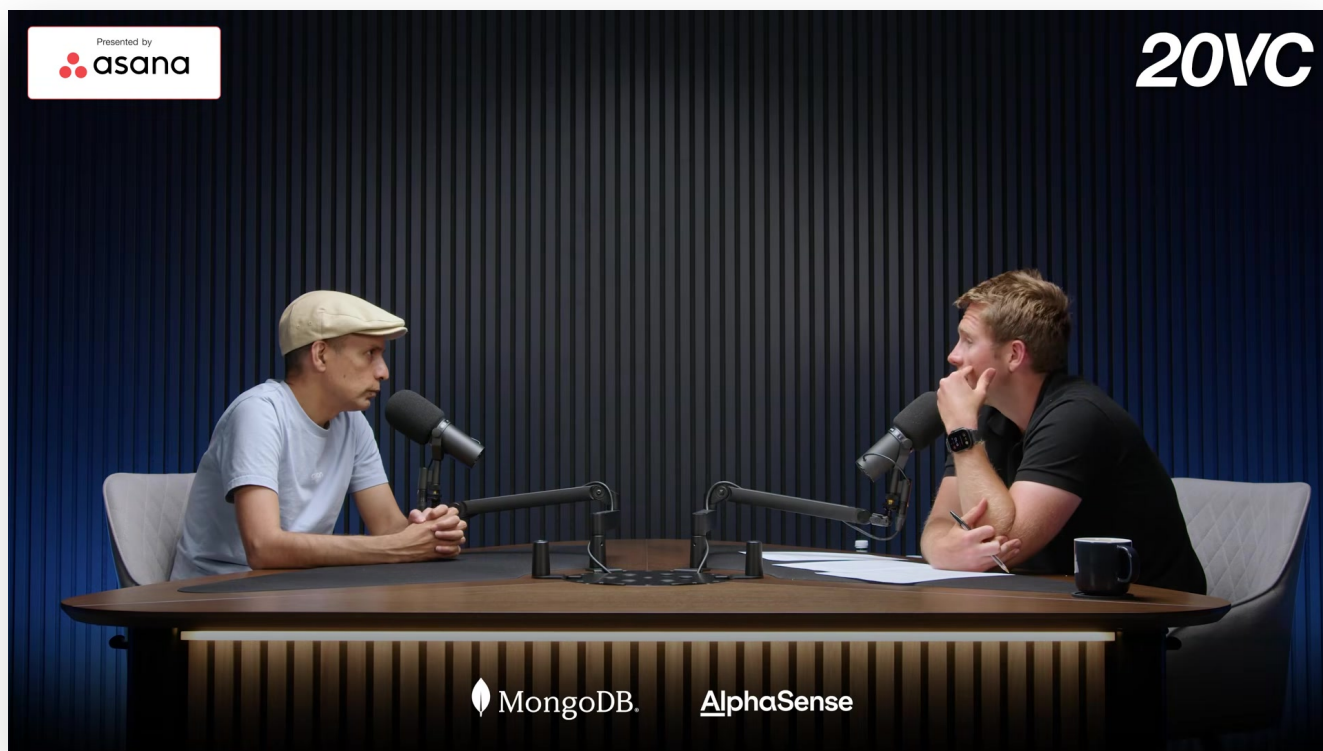
他说，在这种模式下，AI 会变得非常慢。它要花大量时间，先把完成工作所需的基础信息组装出来；同时成本也会非常高，因为大部分 token 都消耗在为某个具体任务拼凑正确上下文这件事上。更糟的是，你还会把 AI 用在一些它其实并不擅长、甚至根本没必要使用它的地方。

所以他说，他们真正主张的是：如果想让 AI 真正发挥效果、真正交付结果，你就必须围绕它进行投入。你必须确保为它提供正确的上下文，这样它才能跑得更快，而且成本更低。

画面说明

这张图切成了双人广角：左边的嘉宾正在说话，双手举在身前比划；右边的主持人托着下巴认真听，桌上还能看到纸张和杯子。这个构图把这段对话的结构直接展示了出来，不再只是嘉宾单向输出，而是“主持人提出定义性问题，嘉宾系统回答”的来回推进。画面里的嘉宾姿态配合他的论点很明确：他在解释 ROI 为什么不是单看模型能力，而是看企业能不能高效地把正确上下文送进 agent；如果只是粗放地把 AI 接到所有系统上，让模型自己暴力搜集材料，结果就会是慢、贵、而且常常用错地方。

19. 别把“用 AI 取代自己”当成目标（27:22–28:52）



场景 19

完整内容

嘉宾说，他们真正讨论的问题是，怎样才能让 AI 真正产生效果、真正交付结果。要做到这一点，你必须围绕 AI 做投入。你得确保给它提供正确的上下文，这样它才能真正与你协同工作，而且速度更快、成本更低。

主持人接着问，所谓“围绕它做投入”到底是什么意思？另外，作为 CEO，如果我们一直鼓励团队成员尽量让 AI 取代自己，哪怕这会浪费很多 token，这样做是不是错的？

嘉宾说，在他看来，把“用 AI 取代你自己”设成目标是错的。首先，这其实是给了 AI 过高的评价。因为当你这样说的时候，你默认 AI 现在已经准备好了，但事实并不是这样。他说，你随便给我举一个例子，说哪个岗位今天真的可以被 AI 取代。比如，你觉得 AI 能取代你的执行助理吗？

主持人回答，就他自己的情况来说，不能。但他也开玩笑说，自己可能是个很难伺候的人。不过对大多数人来说，他认为 AI 确实可以完成这个岗位的大部分工作。

嘉宾回应说，问题也正是在这里。AI 的确可以处理任何一个岗位中的很多事情，但它无法取代最后那一点、最后那种难以言传的部分。

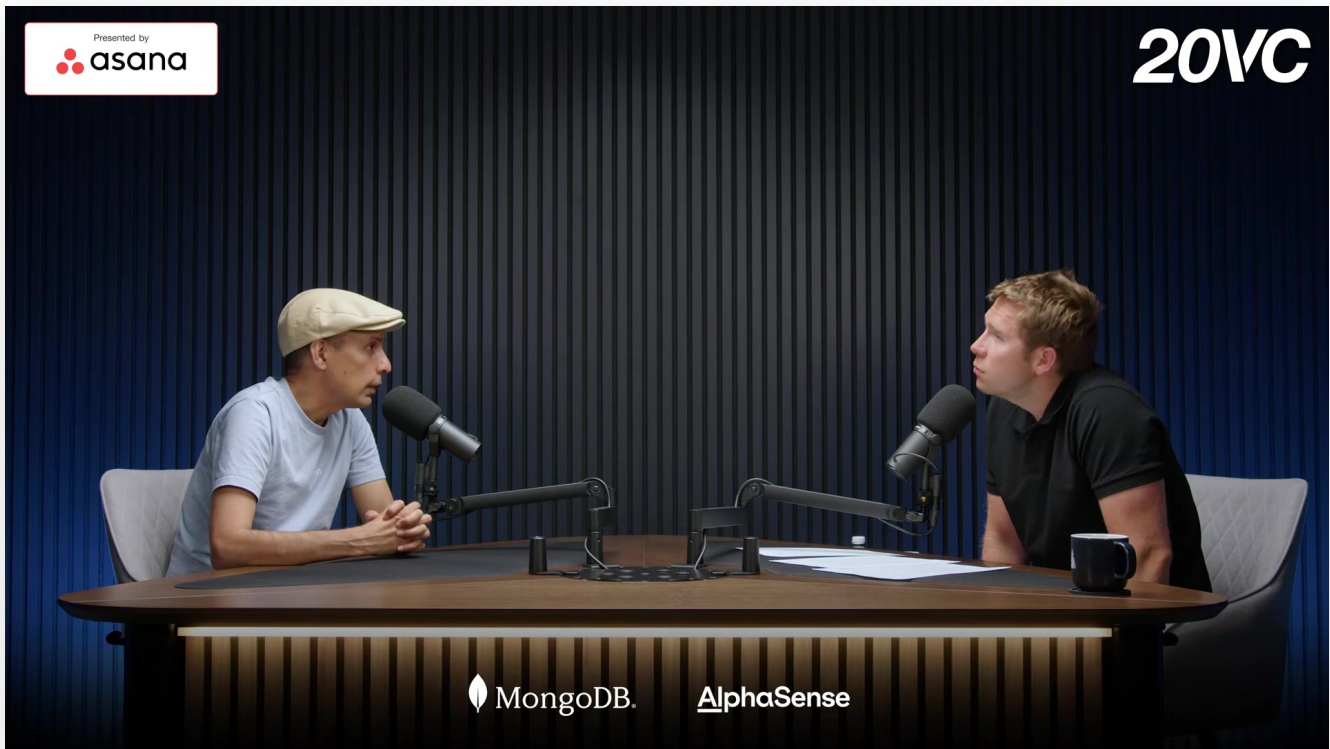
主持人反驳说，但那一点未必重要到无法跨过去。对很多人来说，如果 AI 已经能做 90%，那就够了。比如，给你妻子挑生日礼物这件事，你一年才做一次，不是经常发生的事，而 Claude 可能还不够个性化，没法真正理解你妻子对香水的偏好。所以确实还会剩下一点点做不了的地方，但那个岗位本身还是会被蚕食。

嘉宾说，他不确定，并表示自己会解释为什么。

画面说明

这一帧是两人对坐的广角镜头：左侧戴浅色帽子的嘉宾身体前倾、双手交握，右侧主持人托着下巴在听，画面强调这是一段来回拉扯的辩论，而不是单向输出。对应这一段内容，主持人把问题推到最尖锐的位置，直接问 CEO 是否该鼓励员工“被 AI 取代”；嘉宾则明确反对把“取代自己”当目标，理由是 AI 还没有准备好，它只能覆盖岗位里的大量事务，却触及不到最后那层难以形式化的人类判断与关系处理。主持人进一步把争论推进到“如果 90% 已经够用，会不会仍然造成岗位被蚕食”，因此这一段的核心不是 AI 能不能帮忙，而是“帮助很多”是否等于“足以替代岗位”。

20. 竞争里不会有人满足于九成方案（28:52–30:22）



场景 20

完整内容

嘉宾说，他并不确定主持人的判断，而且他来解释原因。他认为，无论你做什么，你都想成为那个表现最好的人，而他不认为人们会接受一个只有 90% 的解决方案。

主持人接话说，自己并不是因为不在乎成本才这样讲。

嘉宾回答说，这件事其实不是你个人在不在乎成本的问题，而是你必须在工作中和别人竞争。要记住，你的对手也同样拥有你手上的所有 AI 工具；但如果他们在 AI 之上还保留了一个人，那你要怎么和他们竞争？

主持人随后转向组织规模问题，问嘉宾现在公司有多少人。

嘉宾回答说，他们公司现在已经有一千多人了。

主持人接着问，五年之后你觉得会有多少人？

嘉宾说，希望到时候能有 5,000 人。

主持人很惊讶，似乎在说，所以你并不认同缩减团队这条路。

嘉宾说，他们是要继续增长的。

主持人接着表示，但这其实很反常，因为他接触到世界上最大的那些 CEO，而他们每一个都在缩减团队，而且每一个都在说同样的话。

嘉宾直接说，他完全不相信这种逻辑。

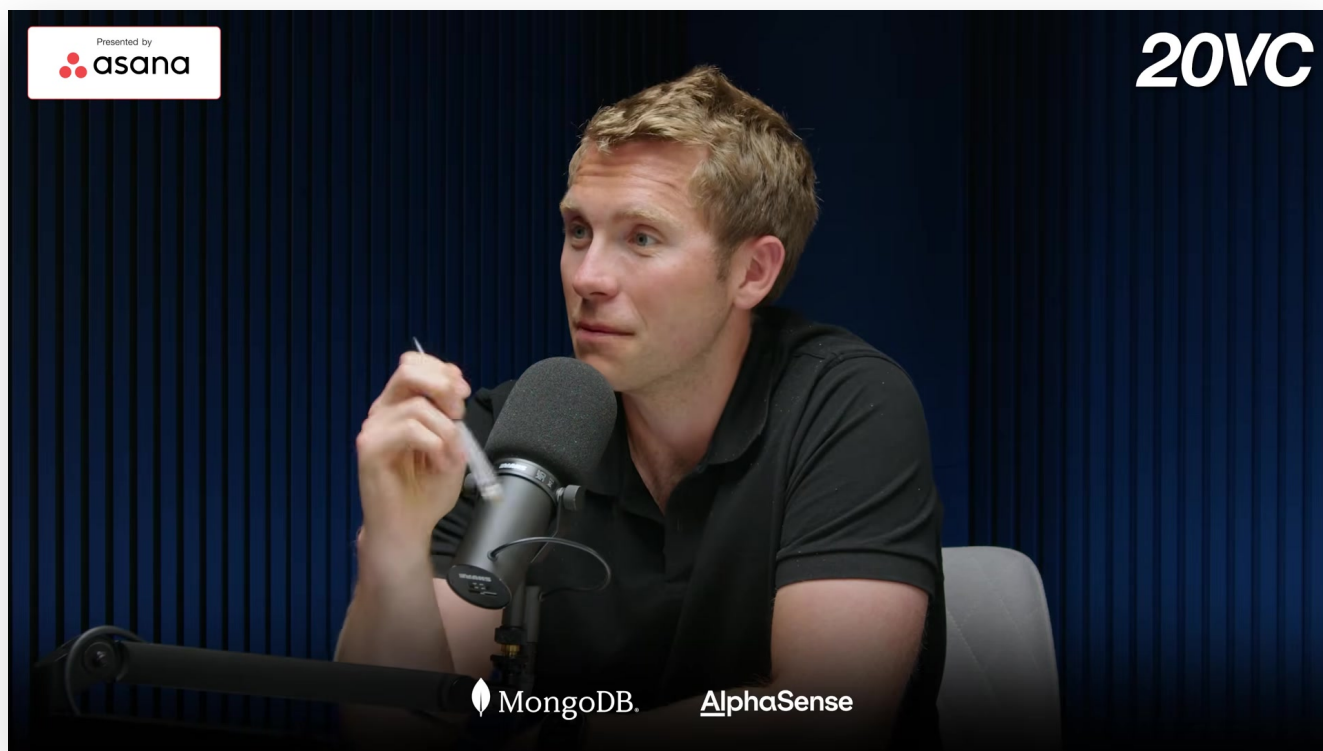
主持人追问，为什么？为什么会这样？

嘉宾说，可以先从逻辑上看这个问题。假设有两家互相竞争的公司，比如可口可乐和百事，或者任何两家正面对抗的公司。其中一家公司决定缩编，另一家公司仍然保留更多员工。两家公司都可以完整地使用同样的 AI 工具和技术。这样一来，问题就变成了：如果你想完成同样数量的工作，并且你相信自己用更少的人也能完成，于是你选择了缩编，那么这到底意味着什么？

画面说明

这张图仍然是双人对谈的中景，但嘉宾左侧正对着麦克风开口，主持人身体前探、认真盯着他，画面重点明显落在嘉宾的回应上。对应的论点是，嘉宾把讨论从“AI 能否做掉 90%”转成“竞争对手也有同样的 AI”，因此真正的比较维度不再是单个员工能否被替代，而是当双方都拿到同样工具时，谁还保留更多人力去叠加在 AI 之上。随后主持人用人数追问他的组织观，嘉宾给出“一千多人、希望五年到 5,000 人”的明确数字，直接表明他不接受“AI 时代公司必然收缩”的主流判断。

21. 同样的 AI，加更多人可以做十倍事（30:22–31:54）



场景 21

完整内容

嘉宾继续说，如果你决定缩编，是因为你相信自己用更少的人也能完成同样的工作，那么你的竞争对手其实也可以这么做。但他们也可以选择不只是完成同样的工作量，而是选择把能力往上抬，做出一个好 10 倍的产品，或者生产出 10 倍更多的东西，因为他们手上有更多人。那样的话，他们就会变得更大，他们会赢，他们会击败你。

主持人反驳说，他不认为更多的人就一定意味着更好的产品。

嘉宾回应说，那是另外一个问题。

主持人接着展开自己的逻辑：如果他可以砍掉 headcount，然后把省下来的钱用来购买最好的 frontier models、最好的技术，让自己的“100x 工程师”拿到最强工具，那会怎样？而且他还认为，人一多，很多事情反而会变慢。

嘉宾说，但这并不是一个 AI 时代才有的论点，这个论点一直都成立。

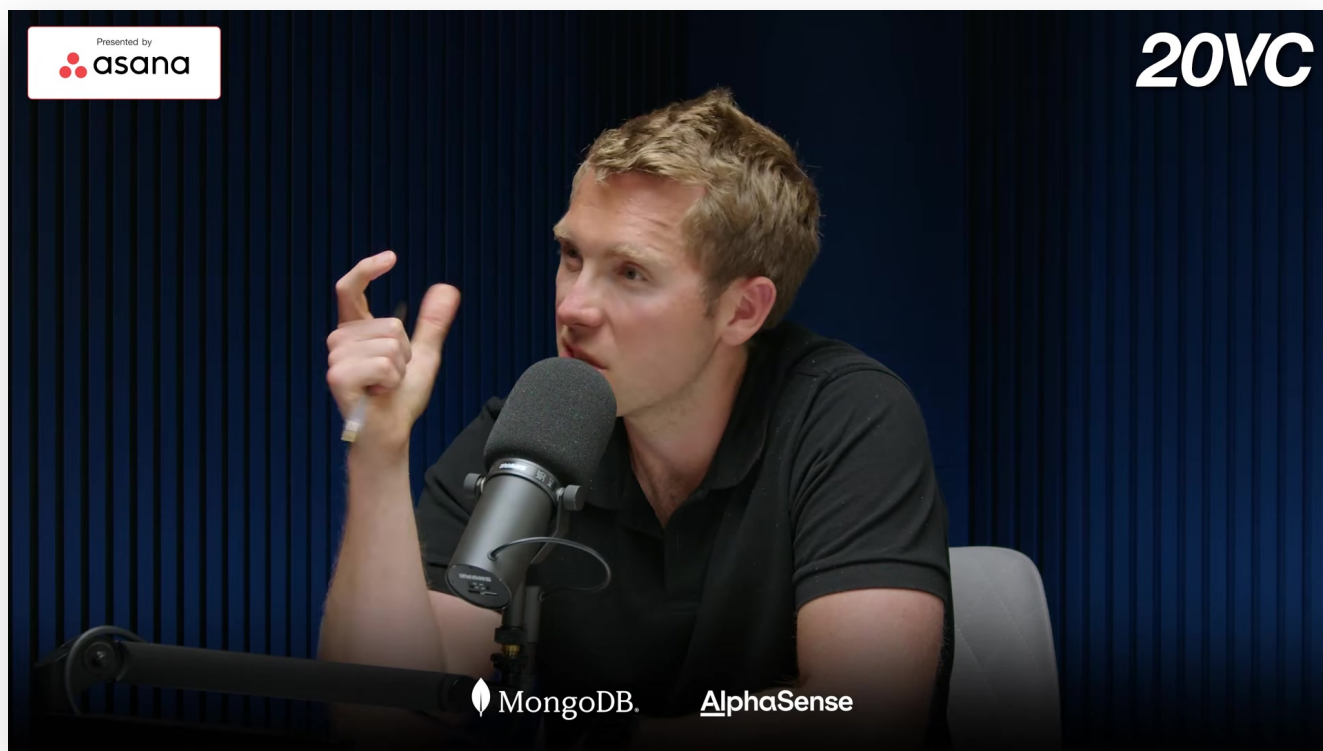
主持人说，没错，但现在叠加 AI 这个因素后，情况不同了。既然你现在能交付更多，那问题就在于，当你有更多人的时候，你也知道会发生什么：他们往往只会竖起更多障碍，阻碍产品真正发布出去。

主持人又补充说，就算先不谈 AI，在那之前，疫情后很多公司也都觉得自己变臃肿了，于是裁掉了 15% 到 20% 的员工，而几乎每一个 CEO 出来都说，裁员之后他们的推进速度反而快了 20%。

画面说明

这一帧切到主持人的近景，他手里拿着笔，对着麦克风发问，说明此时主导话题推进的是主持人。内容上，嘉宾刚刚提出一个很强的竞争假设：在同样拥有 AI 的前提下，不缩编的一方可以把多出的人力用在“做 10 倍更好的产品”或“产出 10 倍更多东西”上，所以会赢。主持人随即把论战拉回自己熟悉的管理经验，强调更多人未必带来更好产品，反而可能拖慢组织，还举出后疫情时期许多公司裁掉 15% 到 20% 人员、速度却快了 20% 的例子。画面上的近景与他手势中的笔，都强化了这段是主持人集中抛出反论证的时刻。

22. 人是资产，不是默认该被减掉的成本（31:54–33:26）



场景 22

完整内容

嘉宾接着回应说，没错，很多公司确实都谈过主持人说的那种情况。这个论点一直都存在：团队到了一定规模之后，就会开始彼此拖慢，人在组织里确实会产生这种摩擦，他自己也相信这一点。

但他接着强调，归根结底，人同样是公司的资产，你必须能够把这些人正确地部署到合适的项目上。所以他并不认为，世界上最伟大的公司最终会是那种只有 100 个人的公司。

他还补充说，看看那些模型公司本身就知道了。既然大家都在讨论 AI，那为什么这些模型公司自己还在如此激进地招聘？

主持人随后提出另一个角度：你难道不觉得，最好的人才想和最好的技术一起工作吗？他猜测，世界上最大的公司在技术上的支出占比，可能会从现在的 8% 到 12%，提高到 16% 到 20%。

主持人继续说，接下来也许会看到 headcount 下降，但 technology spend 上升，而最好的人才更愿意去那些拥有最好工具、最好设备的地方。

嘉宾说，他对这一点也不确定。因为在他看来，技术本来就不应该越来越贵。首先，他反问主持人：你是否承认，当前这项技术的定价，相对于它实际交付的价值来说，是荒谬的？

主持人回答说，这完全取决于它具体在做什么。就像 Cursor 或其他开发工具，他一点也不觉得贵，恰恰相反，他认为它们依然被严重低估了。以 Marc Benioff 在 Anthropic 上花的 3 亿美元为例，那不过只占开发者薪资的 3.7%，他觉得这个比例其实相对很小。

画面说明

画面仍然是主持人的近景，但他的表情更像是在持续追问和设定假说，这与这段内容里他不断抬出“技术支出占比从 8%-12% 升到 16%-20%”的判断相吻合。论证上，嘉宾先承认大团队会互相拖慢，却立即加上关键限定：人本身也是资产，关键不是先砍人，而是能否把人放到正确项目里；他甚至反问，既然 AI 这么强，为什么模型公司自己还在激进招聘。主持人则把视角切到资源配置，主张未来可能是“更少 headcount、更高技术支出”，并用 3 亿美元仅占开发者薪资 3.7% 的数字，为“优质 AI 工具并不贵”做铺垫。这个镜头虽然拍的是主持人，但实际承载的是双方从“团队规模”转入“技术成本结构”的转折点。

23. 自动化做了九成五，账却未必划算（33:26–34:57）



场景 23

完整内容

嘉宾回应说，他会说这其实是荒谬地昂贵。他举了一个例子：他们做了一个很酷的工程分诊 agent。原本他们有一个 15 人的 on-call 团队，这个团队的工作就是去分诊每一个线上生产事故，比如系统告警、各种出问题的状况。

他们后来构建了这样一个 agent，现在这个 agent 已经能自动处理其中大约 95% 的问题。但即便如此，它运行出来的成本仍然很值得怀疑。也就是说，你会问，它真的比人更有效吗？真的比人更高效吗？

嘉宾说，他们当时在这个 agent 上每个月要花 100 万美元。而且那已经比原来人工团队的成本还高。

主持人立刻打断，重复确认说，每个月 100 万美元？

嘉宾说，是的。

主持人开玩笑说，你这是在买克里斯蒂亚诺·罗纳尔多吗？你到底在干什么？

嘉宾回答说，不，他的意思是，这类成本确实非常贵。

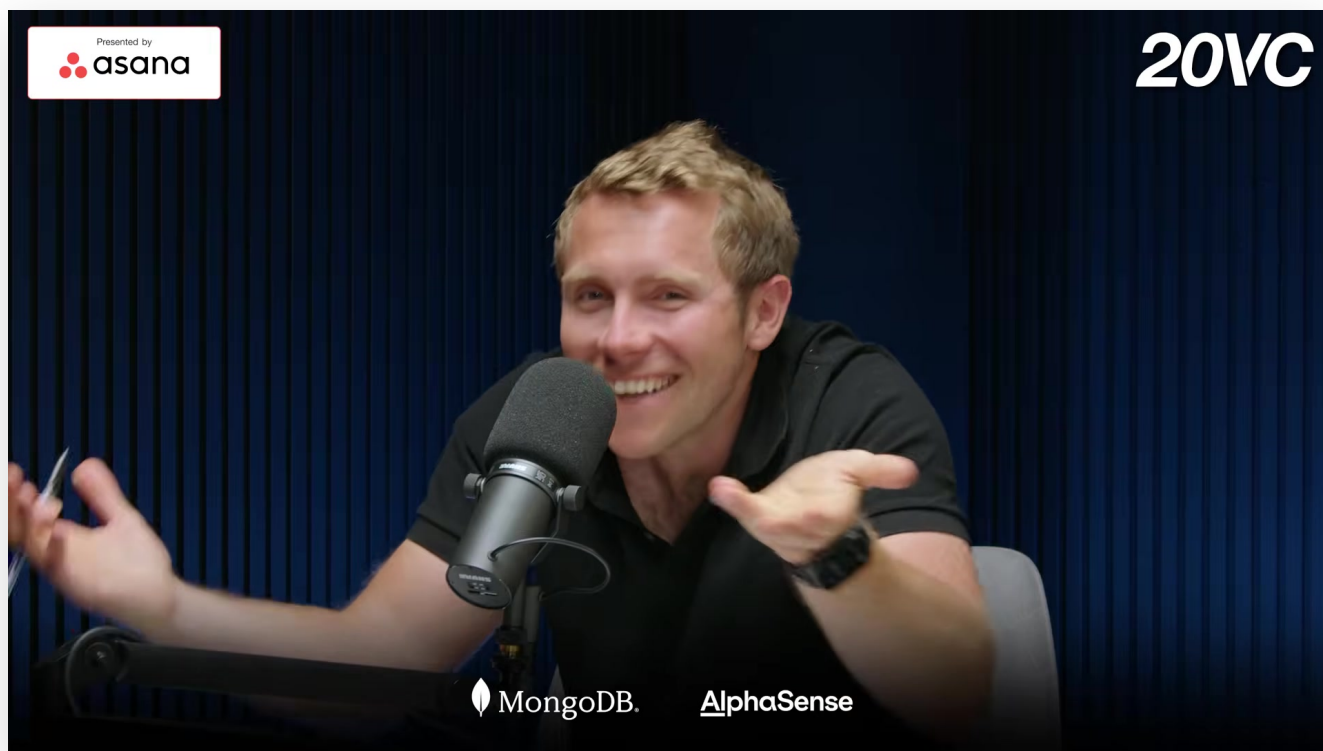
主持人又把话题拉回前面的成本比例问题。他说，等等，让我回到刚才那件事。因为他在节目里经常讨论这个问题，所以嘉宾其实让他变得更聪明了。你的意思是，把开发者薪资的 3.8% 花在这些工具上，就已经算很多了；如果你真这么认为，那这些模型提供商可就彻底完蛋了。

嘉宾解释说，他真正想表达的是，3.8% 这个数字单独看起来其实一点都不高。换个角度看，确实如此。

画面说明

这一帧切到嘉宾近景，他戴着浅色帽子，双手在胸前比划，身体前倾地对着麦克风说话，说明这里是在给出具体案例支撑自己的成本观点。内容上，他不再抽象讨论“AI 值不值”，而是给出一个极具体的工程案例：15 人 on-call 团队、agent 自动处理 95% 问题、但每月成本高达 100 万美元，甚至高于原本人工成本。主持人的惊呼和“你是在买 C 罗吗”这种玩笑，恰好把这个数字的冲击感放大了。于是这一段的关键不是 AI 能不能做事，而是即使它做到了九成五自动化，现阶段的经济账也未必成立。

24. 技术应该越来越便宜，不该越来越像人工（34:57–36:29）



场景 24

完整内容

嘉宾接着说，但他也知道，光是看开源的发展就已经能看到，你完全可以用十分之一的成本完成同样的工作量，这是第一点。

第二点，从他能记事以来，历史上我们从来没有把技术成本和人工成本放在同一句话里比较过。而现在这是第一次，大家真的在说：我宁愿少一些人类员工，多一些 token。他觉得这不是技术应有的运作方式。模型本来就应该越来越便宜，技术也应该越来越可负担。

主持人随后打断并表示抱歉，但他觉得这件事对他来说非常有趣。因为嘉宾是 Glean 和 Rubrik 的联合创始人，而他自己不过是一个播客主持人。但在他看来，这恰恰就是技术存在的意义：agent 变得更主动，拥有判断，能够做决定。既然如此，它们当然应该和人工放在同一句话里讨论，因为它们正在替代过去那些本来要花在人力上的开销。

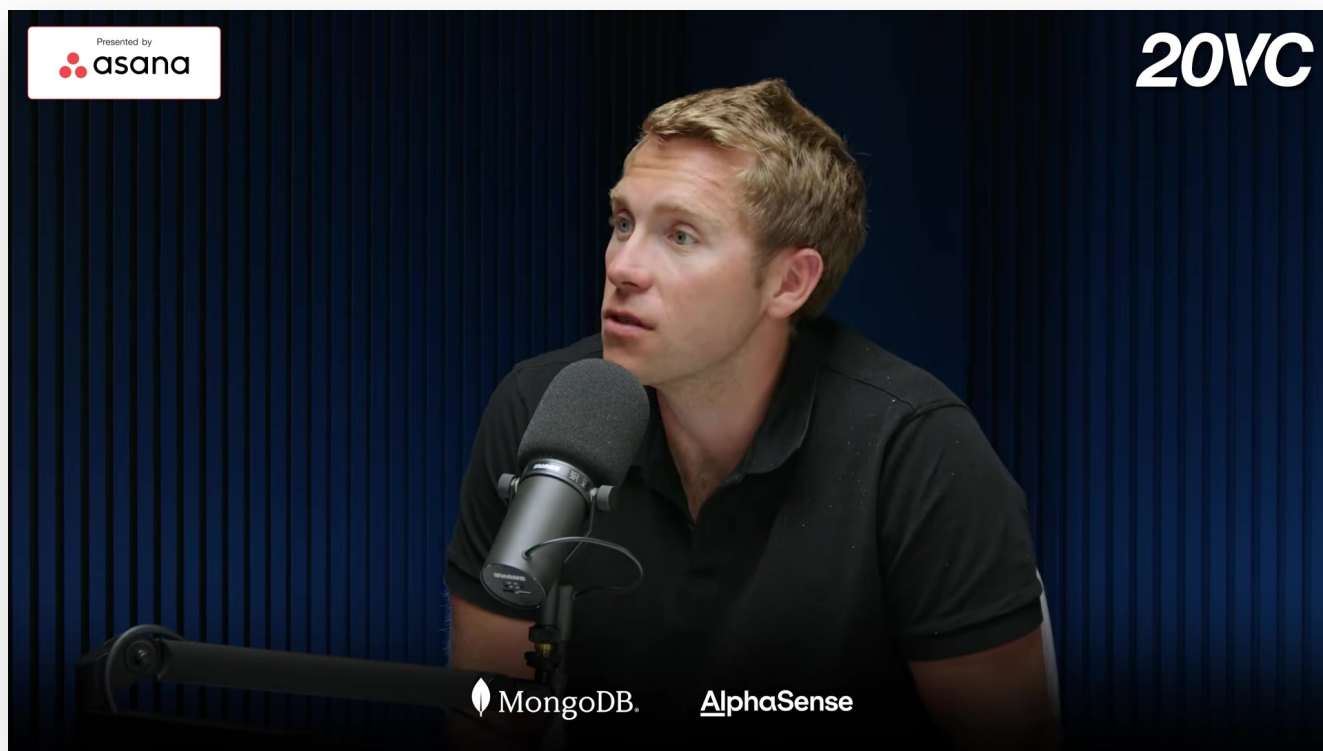
嘉宾回应说，好的技术总会想办法把技术本身做得非常非常便宜，这一次也会一样。

主持人说，这正是他的信念。你会看到推理成本以数量级的方式下降。不过他也认为，过去 6 到 9 个月里其实出现了一件很奇怪的事：几乎每一个模型都提高了自己的单 token 价格。而如果你把时间往回拨 15 个月，那个时候所有人都以为……

画面说明

这一帧是主持人的近景，他张开双手、带着明显的笑意和反问姿态，视觉上很像在说“这不正是技术该做的事吗”。这和本段论证完全一致：嘉宾坚持技术应该像历史上的其他技术一样持续降价，甚至指出开源已经能把同样工作做到十分之一成本，也强调把“少一点人、多一点 token”当成常态，本身就违背了他对技术演进的理解。主持人则明确站在另一边，认为 agent 既然能主动行动、形成判断、做出决策，就应当被视为正在替代人工的技术开支，因此技术成本当然应该与劳动力成本同框比较。画面中的夸张手势与笑着反驳的表情，正好对应这段关于“技术究竟应不应该像人工一样被计价”的正面冲突；而结尾则停在主持人补充的市场观察上，即最近 6 到 9 个月模型单 token 价格反而上升，与大家 15 个月前预期的降价路径形成反差。

25. AI 成本、团队规模与放开试验期 (36:29–38:01)



场景 25

完整内容

过去 15 个月里，大家原本都以为 token 的单价会像之前那样继续一路下降，所以现在这个变化到底是怎么发生的，我们其实也说不清楚，这本身也是一个比较特殊的阶段。

对方接话说，这些公司在上市前需要证明自己是好生意，这就是事情会变成现在这样的原因。他还补了一句：有些话你能说，我不能说。

回答是，我的下注方向仍然是 AI 会比今天便宜得多得多。如果 AI 的成本真的比今天低很多，那么那些本来就在亏损、却在某种程度上支撑着我们整个全球经济的企业，会面临非常大的威胁。

主持人接着确认，所以你的观点还是没变。然后他换了个角度问：这就很有意思了，所以你并不期待未来工程团队会变小，是吗？

回答是，我认为人均生产力会大幅上升，但与此同时，为了赚到同样规模的收入，市场对产品的要求也会一起提高。未来你必须做出一个好 10 倍的产品，不幸的是，现实就是这样。

随后话题转到企业内部的 token 开支。主持人问，你们内部在考虑 token 花费时，作为一个团队是怎么坐下来想这件事的？你和 CFO 一起是如何做 token 预算决策的？

回答是，我觉得我们大概做了多数公司都会做的事，那就是什么都没做。因为我们正处在一个先让大家自己摸索这项技术到底能做什么的阶段。

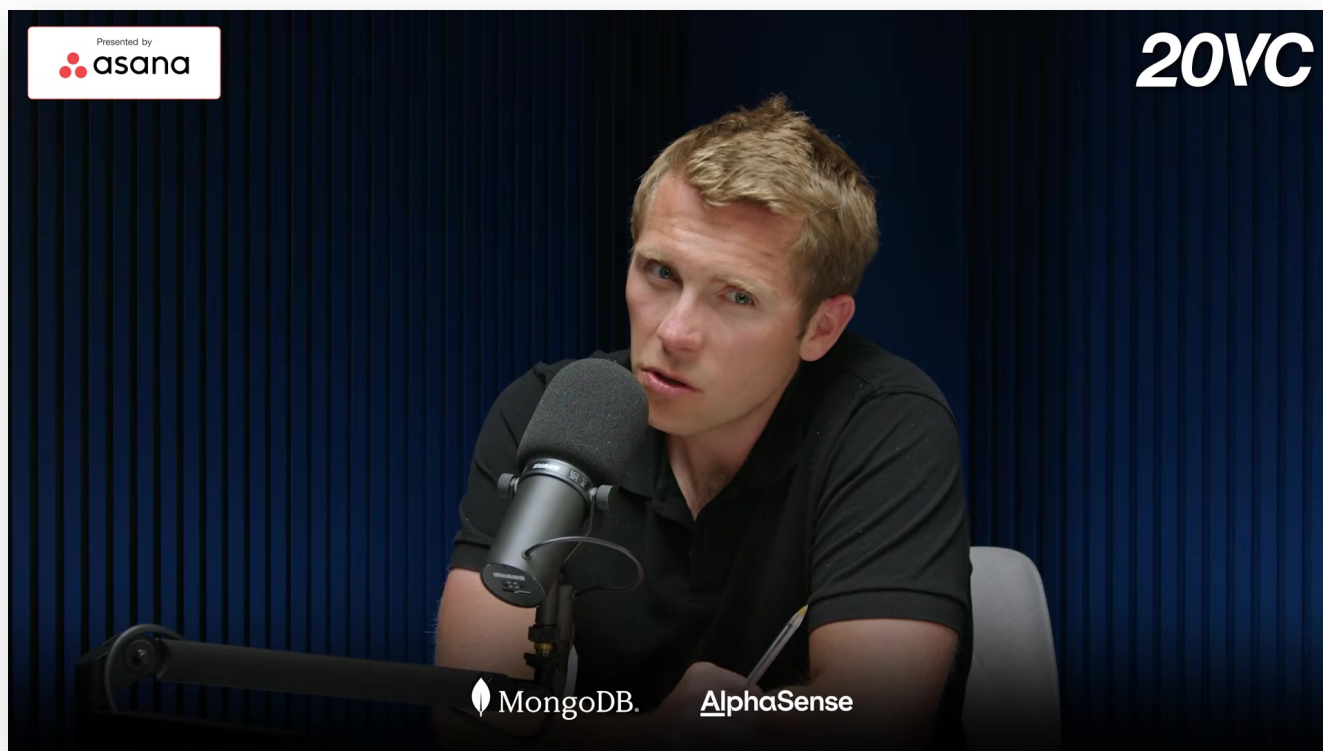
主持人追问，那你们后来看到了什么？大家都疯了一样地用吗，还是大家根本没采用，到底发生了什么？

回答开头是，公司里存在一种幂律分布，而且在我们的客户那里也是一样。

画面说明

这一帧是主持人的近景，他穿黑色 Polo，身体前倾对着麦克风，像是在持续追问并把话题往经营细节上压。这个时间段里，画面对应的论点也确实从宏观判断落到公司管理：先讨论 token 价格原本被普遍预期会继续下跌、而如果 AI 进一步大幅降价，会威胁那些依赖亏损运转的生意；再讨论即使人均效率暴涨，团队未必缩小，因为同样收入目标会倒逼产品必须做到 10 倍更好；最后转入企业内部怎样看 token 预算，而受访者的答案并不是精细管控，而是先放手让员工探索。这种主持人专注发问的画面，和这一段层层推进、从市场成本讲到组织决策的对谈节奏是一致的。

26. Token 使用幂律与基础 AI 普及 (38:01–39:34)



场景 26

完整内容

我们的公司里是这样，我们所有客户那里也一样，你会看到一种幂律分布。有人每个月在 token 上花 1 万美元或者 1.5 万美元，但也有人每个月只花 20 美元。

不过有一点很有意思，就是每个人都在某种程度上接受了 AI。比如，基本上所有人都在使用最基础的那类功能。正如我前面提到过的，今天全球 AI 的第一大应用场景，其实是信息查找和问答。每个人都在做这件事。所以无论是我们团队内部，还是我们的客户，大家都在做这类事。每个人都在提问，每个人都在拿到一些基础的摘要、信息综合和整合结果。

但更高级的 use case 只局限在大概 5% 的员工群体里。

主持人接着问，作为领导者，你会不会主动推动公司尽可能激进地使用 AI？他说我们之前在 Palo Alto 和 Nikesh 聊过，他那边每周都会开领导层会议，做类似 show and tell 的环节，每个人都得站起来展示自己那一周用 AI 做了什么，它是如何替代自己原来的工作、改善工作方式，或者带来了别的

什么变化。你们会不会做类似的事？

回答是，这其实是个很好的主意。我确实想过这么做。我们以前限制过那种 token 使用冲榜的 dashboard，而我一直觉得那不是对的做法。我不认为应该只是去奖励那些消耗更多 token 的人。我当时的感觉是，我们其实不需要这么做。毕竟我们自己本来就是一家原生 AI 公司，大家本来就已经受过足够多的教育了，他们会在需要的时候去使用 AI。

不过，讲到高管层面，比如说……

画面说明

这张图还是主持人的特写，但他的视线更直接、姿态更压低，像是在追问一个运营层面的管理问题。这和这段内容高度匹配：受访者先给出清晰的使用分层，说明企业内部和客户侧都出现了明显的幂律分布，有人每月烧掉 1 万到 1.5 万美元 token，也有人只花 20 美元；再强调基础问答、搜索、摘要已经几乎全民普及，而高级场景只覆盖约 5% 员工。随后主持人把问题推进到管理动作本身，询问领导者是否要用周会、展示机制去强推采用率。画面中的主持人正是这一段驱动讨论方向的人，而整段论证的重心，是“AI 已普及，但深度使用并不均匀，领导者该不该用机制去加速它”。

27. 用案例鼓励 AI 与招聘难度变化 (39:34–41:05)



场景 27

完整内容

我的意思是，在高管层面，分享成功案例当然有价值。我们并没有要求每一个人、每一个星期都必须这么做，但我们确实会做展示。比如在我们的 town hall 里，我们总会让大家分享这些成功案例。几乎每一次 town hall 都会有一个专门的环节，用来讲各个团队正在使用哪些新的 AI agents，大家是如何借此以不同方式完成工作的。

主持人接着问，那么说到你手下的高管和你接触到的人才市场，情况怎么样？

回答是，我觉得招聘从来没有像现在这么难过。

主持人追问，在今天这个环境里招聘到底有多难？尤其是像我们前面说的，那些最大的模型提供商开出的薪资高得惊人，已经到了我们以前从没见过的水平。

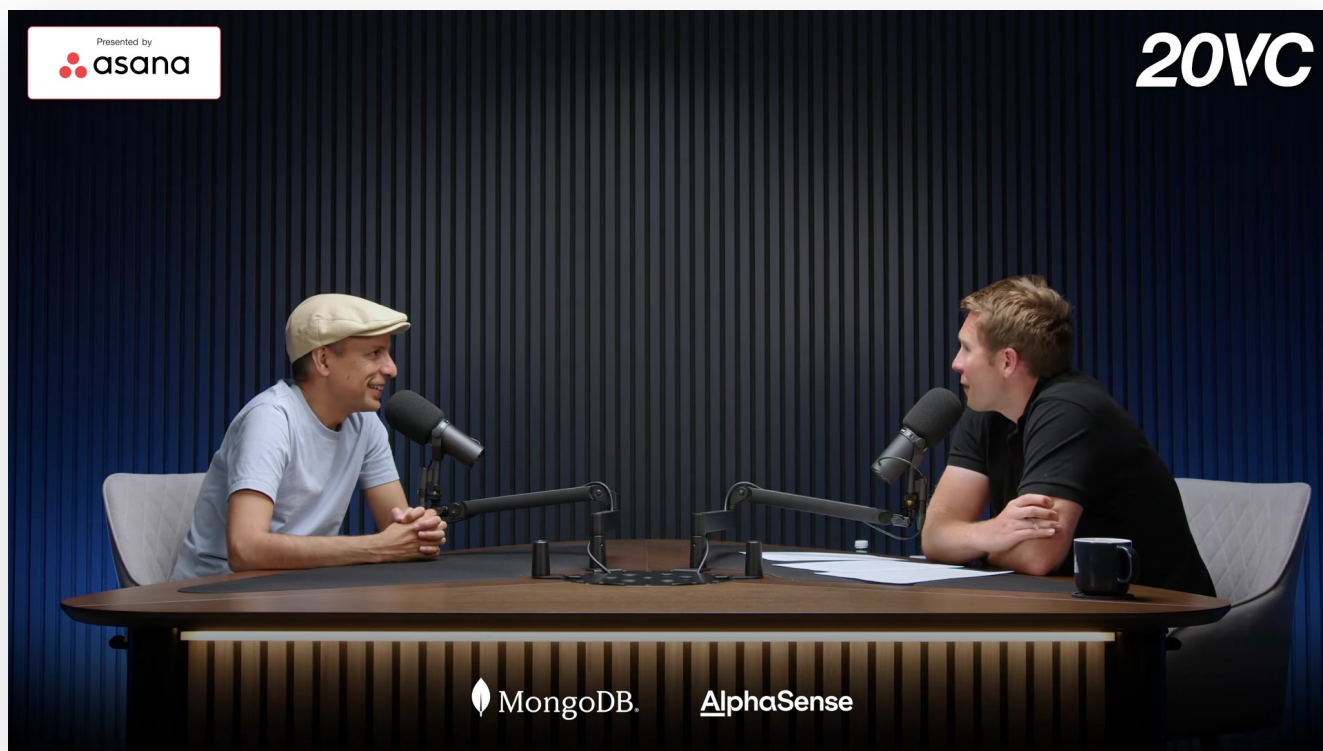
回答是，嗯，我其实会说，先把最近两三个月单独放在一边不谈。如果不看这最近两三个月，我会说，相比 SaaS 最疯狂的高峰期，招聘其实一度是在变容易的。为什么会这样？

回答是，因为很多公司都没有再继续扩张 headcount。你去看那些最大的科技人才雇主，很多公司实际上都没有继续增长，很多甚至一直在裁员。就拿 Meta 来说吧，几乎每年都会有比较显著的裁员。我不知道它的总 headcount 现在具体是多少，但我猜大概率已经低于 2021 或 2022 年高点时的水平。所以实际上，市场上可用的人才是比以前更多的。

画面说明

这一帧切到了受访者的近景，他穿浅蓝色 T 恤、戴浅色帽子，靠近麦克风，像是在给出自己的判断而不是单纯回应细节。画面里由他主讲，也正对应这段内容的核心：他先解释公司并不会用强制性的每周制度推动所有人汇报 AI 使用，而是通过 town hall 固定展示团队如何用新的 AI agents 改变工作方式；随后话题转向招聘，他给出一个带时间切分的、更细的判断，即如果先把最近两三个月单独排除，整体招聘其实曾比 SaaS 泡沫高峰期更容易，因为大型科技公司停止扩编、持续裁员，释放了更多人才。这个近景强化了“受访者在做结构化判断”的感觉，也承接了从内部管理转向外部人才市场的论证。

28. 顶级 AI 人才推高种子轮需求 (41:05–42:39)



场景 28

完整内容

所以从整体市场来看，可用人才其实比以前更多。但如果你现在开始具体谈 AI 人才、ML 人才，尤其是顶尖人才，那情况就完全不同了。顶尖的人比以往任何时候都更被争抢，而且薪酬水平也彻底变了。

而且不仅仅是模型公司把价格抬高了，连创业公司也是一样。因为创业公司为了竞争人才，也在给这些人开出非常高的待遇。甚至今天的创业公司也得付很多钱。我们就得这样做。

主持人接话说，对，我们也必须这么做，因为这些人的备选选项太大了。

接着他又说，而且如果一个顶尖开发者要花 30 万、40 万，甚至 50 万美元，

对方回应，是的。

主持人继续说，那一个 200 万美元的种子轮基本上根本不够用。

接着他又补了一种创始人的现实情况：假设创始人自己都不拿工资，但如果我要雇 4 个人，

回答是，那我就需要 600 万美元。

主持人说，没错。

然后主持人问，你觉得创始人是不是应该把种子轮做大？

回答是，我觉得更好。我的一贯偏好一直是，在一轮融资里，只要你能做到，就尽量从一开始多融一些。

主持人又问，你们哪一轮融资让你感觉估值最高、最“贵”？

回答是，首先，其实除了公司最早的那一轮，我们从来没有真正主动出去融资。一般都是有人主动找上门来，关系经过一段时间慢慢建立，最后基本上就顺势变成一个默认结果，也就是他们会成为投钱进来的人。我会说，我们的 C 轮大概是最……

画面说明

这张图是两个人的全景对谈：受访者和主持人分坐桌子两侧，身体都前倾，明显处在一来一回、节奏很快的讨论里。全景很适合这一段，因为这段不是单向陈述，而是围绕“高薪人才如何改变创业融资逻辑”的共同推演。受访者先把整体科技人才市场和顶级 AI/ML 人才市场区分开，说前者供给更松，后者却被前所未有地争抢，连创业公司都被迫抬高薪资；主持人再把抽象判断换算成具体数字，提出 30 万、40 万、50 万美元一个优秀开发者的成本，以及 200 万美元种子轮不够、雇 4 个人就要 600 万美元的现实。画面里两人面对面交锋，正对应这段从人才竞争一路推到融资规模的完整论证。

29. C 轮高估值与投资品牌效应（42:39–44:10）



场景 29

完整内容

我会说，我们的 C 轮大概是感觉最“贵”的一轮。你可以说它的价格最高，因为当时我们的业务其实非常少，肯定低于 200 万或 300 万美元，也许是 500 万美元，我现在记不清精确数字了，但估值已经超过了 10 亿美元。所以那确实是非常极端的一轮。不过我想，我们也是有什么就接什么。

主持人说，这很不可思议。那你在做这件事的时候，会不会担心自己之后能不能长进这个估值？还是你只是埋头往前干，觉得这是一个高价格、低稀释的好交易？

回答是，我们当时更看重的是，这会向潜在员工传递一个信号。比起别的，它更像是一种要对外做出的声明。我们想让市场明白，我们正在打造某种特别的东西，而这样的融资本身会给我们带来那种外部验证。

主持人问，员工真的会在乎你的投资人是谁吗？

回答是，绝对会在乎。我的意思是，

主持人接着说，很多创始人会说，最优秀的人根本不在乎这些，他们是为了使命来的。我总会说，我向你保证，只要你的投资人里有 Coatue，或者有 DST，或者有 Sequoia，优秀候选人突然就会更愿意跟你聊很多。

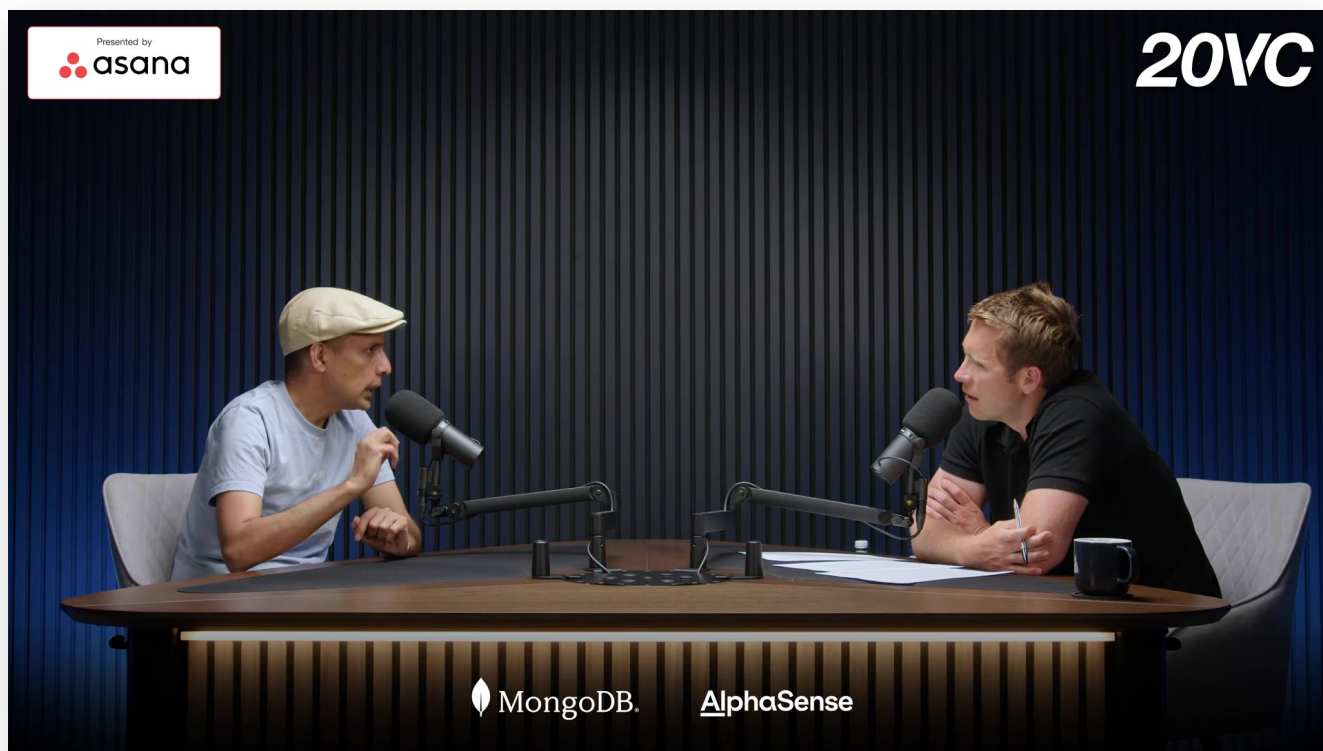
回答是，对，投资人的声誉会直接影响你的声誉。

主持人随后换了个更个人化的问题：如果看今天，相比过去 12 个月，你在哪件事上改变主意最多？一直以来，我个人的风格可能有点过于克制了，已经不太适合现在这种竞争策略了。我的团队会给我这样的反馈……

画面说明

这一帧回到受访者近景，他一边对着麦克风说话，一边用手势辅助，像是在解释一件自己亲历过、而且带有策略意味的融资决定。这与本段内容完全一致：他具体回顾 Glean 的 C 轮，指出当时业务规模还很小，也许只有 200 万到 500 万美元级别，却拿到了超过 10 亿美元估值，因此那一轮“很贵”；但他们接受这件事，不只是为了低稀释，更是为了向潜在员工和市场发出“我们在做特别的事”的信号。接下来两人还明确讨论了投资人品牌对招聘的直接作用，甚至点名 Coatue、DST、Sequoia 会显著提升候选人的兴趣。画面中的受访者是在讲自己的动机和经验，因此能自然承接这一段关于估值、信号和声誉传导的完整论证。

30. 纪律经营与 AI 市场抢地盘（44:10–45:42）



场景 30

完整内容

我的团队会给我这样的反馈：我们现在太保守了。我们在努力确保资本能够撑得更久，但在这种思维模式下，我们可能会错过这一轮抢地盘。

所以我自己也在感受到那种压力，逼着我去改变，改变我对“该怎么花钱、该怎么投资”的看法。但与此同时，我一直有一个根本性的信念，那就是一家企业始终是建立在纪律之上的。你必须为产品收费，它必须真正为客户创造价值。你在营销上每投入 1 美元，背后就必须有合理的回报。你不能假设自己只要不断融资，就能把所有那些本来不存在的东西都补出来。

主持人问，当你看到像 Uber 这样的例子时，你还认同这套看法吗？因为 Uber 似乎证明了，一个糟糕的商业模式也可能随着规模扩大而变成一个好生意。

回答是，对，我说的就是这个。这正是我会感到那种压力的原因，因为它会让我怀疑，也许我原来的思考方式是错的。

主持人追问，你觉得这真的是一场抢地盘吗？

回答是，绝对是。毫无疑问，我们现在就在一场抢地盘里。今天世界上的每一家公司都想要一个像我们这样的产品。要么我们现在就进去，要么未来要进入就会难 10 倍。

主持人接着说，我们前面谈过岗位替代的问题。我们有过一段很有意思的讨论。接下来他转向下一个话题：今天还不存在、但你认为未来会出现的职位是什么？

画面说明

这一帧是两人的全景，受访者在左侧举手强调，主持人在右侧抱臂前倾认真听，桌上还有纸张和杯子，整个姿态像是在讨论一组有张力的经营分歧。内容上，这一段的核心也正是这种张力：受访者一方面承认团队在推动他放弃过度保守，因为如果过分强调资本效率，就可能错过 AI 市场的“land grab”；另一方面，他又坚持企业必须建立在纪律上，产品要收费、要创造价值、营销投入要有回报，不能靠不断融资掩盖基本面。主持人则用 Uber 作为反例，挑战“坏商业模式也能靠规模转正”的现实，逼他承认自己的思路也在被动摇。最后双方在“这是不是一场抢地盘”上达成了明确结论：是，而且现在不进入，未来会难 10 倍。全景画面把这种来回拉扯和最终落点都完整地承住了。

31. 复合型岗位会越来越普遍（45:42–47:14）



场景 31

完整内容

主持人问：你认为现在还不存在、但在未来 3 到 5 年会变得非常常见的工作是什么？

嘉宾回答说，复合型岗位会变得非常普遍。比如说，未来会出现这样一类人：他们能够独立构建产品。嘉宾一时也不知道该怎么给这类人命名，但他们会同时扮演工程师、产品经理、设计师的角色。类似地，在 go-to-market 侧，也会出现能够既卖产品、又能亲自完成产品演示、还能深入讲清楚使用场景的人，而不是像现在这样，把客户经理、解决方案工程师、销售解决方案架构师分得很开。嘉宾认为，未来会看到越来越多岗位走向泛化，而不是继续走向高度专业化。事实上，他甚至一直在自己公司里非常努力地推动这种变化。

主持人随后提出异议，并表示自己是善意地提问：复合型岗位这个说法，听起来恰恰和“维持团队规模”的想法相反，因为如果一个岗位要把四种不同专长都装进一个人身上，这不就是把更多能力塞进同一个角色吗？

嘉宾回应说：这反而正是更小团队的体现。

主持人说：确实如此。

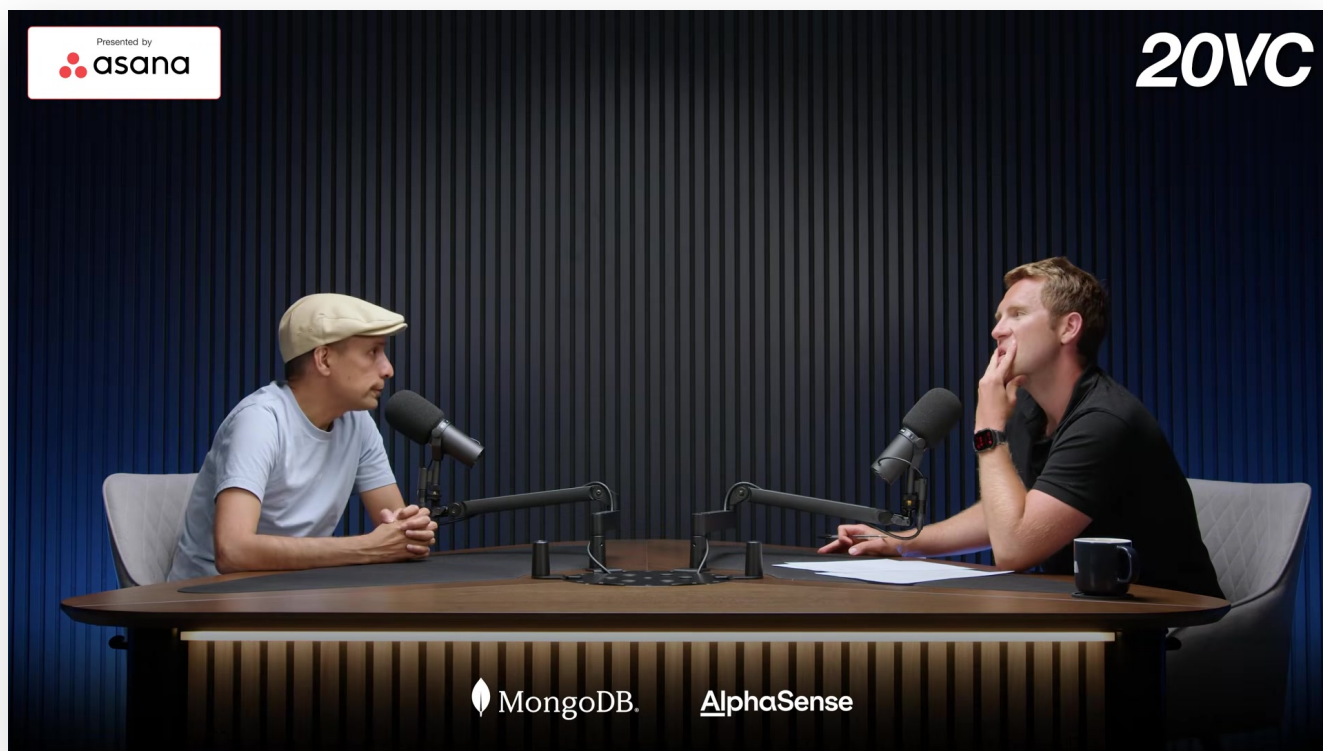
嘉宾接着解释，但正如他之前说过的，未来为了拿到和今天一样的客户收入，你必须做 10 倍的工作。也就是说，你必须用更小的团队，去交付过去同样规模、甚至更多的工作。大家只是被迫要做得更多。

主持人表示明白了，并准备进入下一个问题。

画面说明

这一帧是嘉宾的近景，他戴着浅色帽子、穿着带 glean 标识的浅蓝色 T 恤，身体前倾，对着麦克风认真解释，双手在胸前比划，画面强化了他在主动定义一个新工作形态。这个时间段的核心论点是：AI 不只是让单点岗位提效，而是会推动岗位边界被打穿，形成能同时覆盖产品、工程、设计，或同时覆盖销售、演示、场景讲解的复合型角色。主持人随即挑战这一判断，认为这似乎意味着把四种专业能力叠加到一个人身上；嘉宾则明确反驳，说这恰恰对应更小团队，因为未来为了获得同样的收入，组织不得不用更少的人完成过去更多的工作。画面里嘉宾持续主导发言，与这段关于“泛化岗位”和“更小团队被迫承担更多产出”的论证完全一致。

32. 哪些现有岗位会被吸收掉（47:14–48:44）



场景 32

完整内容

主持人接着问：那今天已经存在的哪些岗位，未来会不复存在？有哪些工作会让我们回头看时觉得，“天啊，我们以前居然还这么干”？

嘉宾回答说，很多分析师类岗位都会消失，或者说会被吸收掉。比如那种并不是业务思考者的数据分析岗位，他们只是接到一个任务，说“我需要看这部分数据”，然后就去制作特定的仪表盘、配置后端系统。嘉宾认为，这类工作肯定会消失。商业智能这个领域会变得非常不一样，因为业务负责人将能够直接得到自己问题的答案。所以，商业分析师、数据分析师这类岗位就是一个例子。

他又补充说，很多 HR 相关岗位也是如此，比如招聘中的 sourcer。这个角色他认为也一定会被吸收到更完整的全流程招聘岗位里。

随后主持人说，他必须再问最后一个问题，而且因为他们此刻正坐在欧洲，这引出了一个关于“主权”的问题。主持人直言，美国和欧洲说得难听一点，在开源这件事上并没有真正达到应有的水平。

嘉宾先回应了一声“对”。

主持人继续问：你觉得未来会不会出现一个“主权模型”的世界？而且考虑到过去一个月左右我们看到的事情，我们是否必须对自己的模型拥有主权？

嘉宾回答说，对主权模型的需求是强烈的，实际上他甚至会说，这种需求大概在一年前比现在还更强。至少就他的体感而言，现在听到的这类声音比之前少了一些。

画面说明

这一帧拉成双人中景，嘉宾坐在左侧，主持人坐在右侧，主持人一手托着下巴、身体前倾，明显处在追问状态，桌上还能看到纸张和杯子，画面强调了对话从岗位变化转向更宏观的政策与地区问题。这个片段前半段由嘉宾主导，他列举会消失或被吸收的岗位，包括不负责业务思考、只执行报表和后台配置任务的数据分析角色，以及招聘中的 sourcer；后半段则由主持人把讨论切到欧洲语境下的“模型主权”，追问是否需要主权模型，以及过去一个月的事件是否说明这种主权是必要的。双人同框准确对应了这段从具体岗位判断过渡到地缘与技术主权争论的节奏。

33. 主权模型需求是在上升还是降温（48:44–50:14）



场景 33

完整内容

嘉宾接着刚才的话说，他会说一年前相比现在，大家对主权模型的呼声其实更强一些。至少按他的感受，这类讨论现在听到得少了。

他解释说，曾经有一段时间，几乎每个国家都觉得自己可以造出一个模型。那是 AI 还处在更早期阶段的时候。但后来很多国家其实意识到，这条路未必会成为现实，所以他们现在也接受让本国企业去使用 OpenAI、Anthropic，或者其他各种模型。

不过嘉宾也补充，他并不是这个领域的专家，所以他并不确定这个趋势到底是在上升，还是在某种程度上略微下降。

主持人立刻表示不同意。他说，考虑到他们在特朗普政府时期看到的事情，这个趋势毫无疑问是在上升。比如特朗普政府禁止了 Anthropic 的最新模型，这让很多人，尤其是很多欧洲人，更深刻地意识到：他们不能依赖某一个美国个人，因为那个人可能会切断他们获取智能能力的渠道。

嘉宾回应说：是，但结果在哪里呢？

主持人说：那不过是一个月前的事，所以如果要求 3 周之内就出现一个现成的替代模型，那确实很难。

嘉宾继续表示，是的，是的，但即便在那之前，这件事其实也一直没有真正发生。按他的看法，世界上真正做出了美国之外模型的国家，只有中国。

主持人回应说：对。

嘉宾又补充说，也许法国算有一点点，像那位先生那样。

主持人接着追问：这是不是一个激励问题？是不是因为美国缺乏开源社区？

画面说明

这一帧回到嘉宾近景，他面对麦克风，双手摊开，像是在做带保留意见的解释，这和他在这段里不断加上的限定语一致，比如“按我的感觉”“我不是专家”“我不确定趋势是在上升还是下降一点”。他的完整论点是：主权模型的诉求并非线性增强，一年前甚至可能更强；很多国家最初以为自己能造模型，后来接受了让本国企业使用 OpenAI、Anthropic 等外部模型。主持人则明确反驳，认为特朗普政府禁止 Anthropic 最新模型这类事件，尤其让欧洲人意识到不能把智能能力绑定在某个美国个人手里，因此主权需求实际上正在上升。画面里嘉宾主讲的姿态，与这段“谨慎判断趋势”并被主持人强硬挑战的争论相匹配。

34. 美国没有开放模型不是因为缺开源文化（50:14–51:45）



场景 34

完整内容

主持人追问：这是不是一个激励问题？是不是因为美国缺乏开源社区，所以我们在美国没有任何真正有规模、也有实质内容的开源成果？

嘉宾回答说，不是。他认为美国其实有很好的开源社区，只是在很多别的领域里表现得很强。

主持人马上插话：可美国到底有什么开放模型呢？

嘉宾承认：对，这点你说得对，我们确实没有开放模型。但这并不是因为开源作为一种运动、作为一种理念，在美国很弱。事实上它在美国相当强。

嘉宾接着解释，如果只谈模型本身，它们需要大量前期投资，而这在很多方面都不利于开源。大量开源软件最初其实都像 skunkworks 一样，是开发者在几乎没有任何资金支持的情况下自己做出来的，最后仍然做成了东西。但模型没法用这种方式造出来，所以这也是为什么事情自然没有那样发展。

他进一步说，需要的是一些新的技术路径，让“超高投入”不再是必需的。

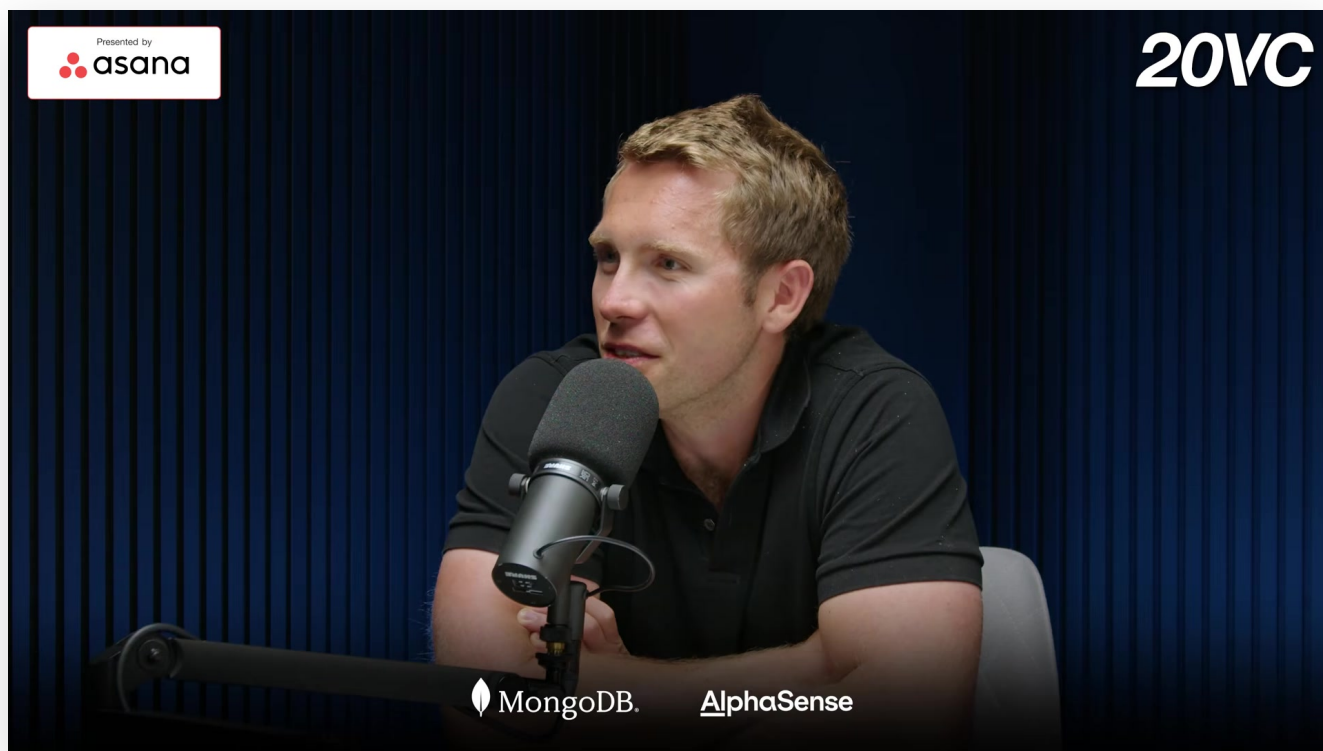
接着主持人把话题又往前推了一步：那你会不会担心现在的局面？他说自己花了很多时间在 OpenRouter 上看模型使用量和流量数据，Anthropic 现在是排名最高的美国模型，但也只是第七名，前六名全都是中国模型。那他们是不是应该干脆放弃原来的那套想法？中国政府在资助排名前六的模型，这件事难道无所谓吗？

嘉宾开始回应：之所以大家会对这些模型感到放心，是因为你实际上可以在一个受控、封闭的环境里运行它们的推理。

画面说明

这一帧切到主持人近景，他穿黑色上衣，双臂交叠靠近桌面，对着麦克风发问，视觉上很像在步步紧逼地质疑对方。这个时间段的论证结构也是如此：主持人先把“美国没有开放模型”归因为“美国缺乏开源社区”，嘉宾明确否认，强调美国缺的不是开源文化，而是适合模型研发的资金结构，因为模型需要巨额前期投入，不能像很多无资金支持的开源软件那样靠零散开发者自发做出来。他随后提出，要依赖能降低超高投入门槛的新技术。主持人又拿 OpenRouter 的使用和流量排序继续施压，指出最强的美国模型也仅列第七，前六都是中国模型，并追问这是否意味着更深的战略问题。画面里主持人的主导发问，与这段连续追击式的质疑完全一致。

35. 监管会不会被用来打压开源 (51:45–53:18)



场景 35

完整内容

嘉宾继续回答说，正因为你可以在那种受控环境里运行这些模型的推理，所以人们会觉得更安心。但他并不认为，美国会对这种趋势完全无动于衷、觉得没问题。现在其实已经有一些很好的工作在推进，目的是在美国真正促进开源和模型研发，也会有一些模型陆续出来。

主持人接着说，另一种可能是：Sam 和 OpenAI 给特朗普 5%，然后特朗普通过监管俘获来保护 Anthropic 和 OpenAI，同时去打击开源。

嘉宾回应说，他希望不会发生这种事，他也怀疑事情会发展到那个地步。

主持人追问说，不好意思，他是在学习中，但 Sam 为什么还会给出那 5% 呢？这不就是一种 quid pro quo 吗？我需要你，你也需要我。

嘉宾回答说，他更相信美国制度本身最终会起到约束作用。而且顺便说一句，他认为现在其实还没到需要去压制开源的时候，因为美国的开源模型目前还太落后了。

主持人继续问：你不觉得 Sam 和 Dario 正坐在那里想，“哇，我们低估了这件事，而这已经成了我们业务的核心威胁”吗？

嘉宾说，他的意思是，他们大概确实会这么想，但他不认为他们能通过监管来解决这个问题。

主持人继续追击：你不觉得 Sam 会直接打电话给特朗普吗？毕竟他能直接联系到他。他可能会说，中国政府正在资助你、也就是我们的最大竞争对手，而且我们不能保证这些模型里没有通向习近平的后门。你得阻止这一切。

画面说明

这一帧中嘉宾再次占据画面中心，手搭在胸前靠近麦克风，表情谨慎，像是在一边承认风险、一边压住结论的强度。这正对应他在这一段里的立场：他承认美国不会对中国开放模型的上升趋势坐视不理，也承认 Sam、Dario 这类公司领导人大概率会把开源视作核心威胁，但他仍然不认为问题能靠监管解决，也不认为当前美国需要先去压制开源，因为美国开源模型还“太落后”。与之相对，主持人的整个论证更尖锐，甚至提出“给特朗普 5%”可能构成 quid pro quo，并设想 Sam 借国家安全叙事推动对中国开放模型设限。嘉宾在画面中的克制姿态，准确对应这段关于监管俘获、政治交换与产业竞争的高强度争论。

36. 美国需要自己的开放模型与多模型世界 (53:18–54:49)



场景 36

完整内容

主持人把自己的设想继续说完：他认为 Sam 可能会对特朗普说，中国政府正在资助你们最大的竞争对手，而且他们无法保证这些模型里没有通向习近平的后门，所以你必须阻止这件事；而作为交换，“我会给你 5%，算是你的辛苦费”。

嘉宾回应说，难道这个论点不也可以反过来理解吗？现在有这么非常好的开源模型，而且它们都来自中国，美国就必须建立自己的能力。美国不能被看成一个在技术上不再创新的国家，所以对美国来说，自己把这件事做出来其实至关重要。

主持人则说，他觉得 Sam 会强调的是：这件事需要数十亿美元和数年时间。特朗普应该“保卫美国”，支持 OpenAI 和 Anthropic，并设置障碍，阻止中国开放模型获得采用，比如征税、禁令之类的办法。

嘉宾回答说，这些措施或许有可能，但美国自己的开源模型会拥有很多顺风。他说，这已经是湾区每个技术人都会讨论、而且普遍承认的问题。有很多动力很强的参与方都希望推动这件事，例如英伟达就在投入很多资源，去促进美国优秀开源模型的发展。

主持人说，他希望他们能成功。

嘉宾表示，当然。一个多模型世界对所有人都很重要。

主持人接着转换节奏，说接下来想和他做一个快速问答：自己说一个很短的陈述，请嘉宾立刻给出反应，这样可以吗？

嘉宾答应了。

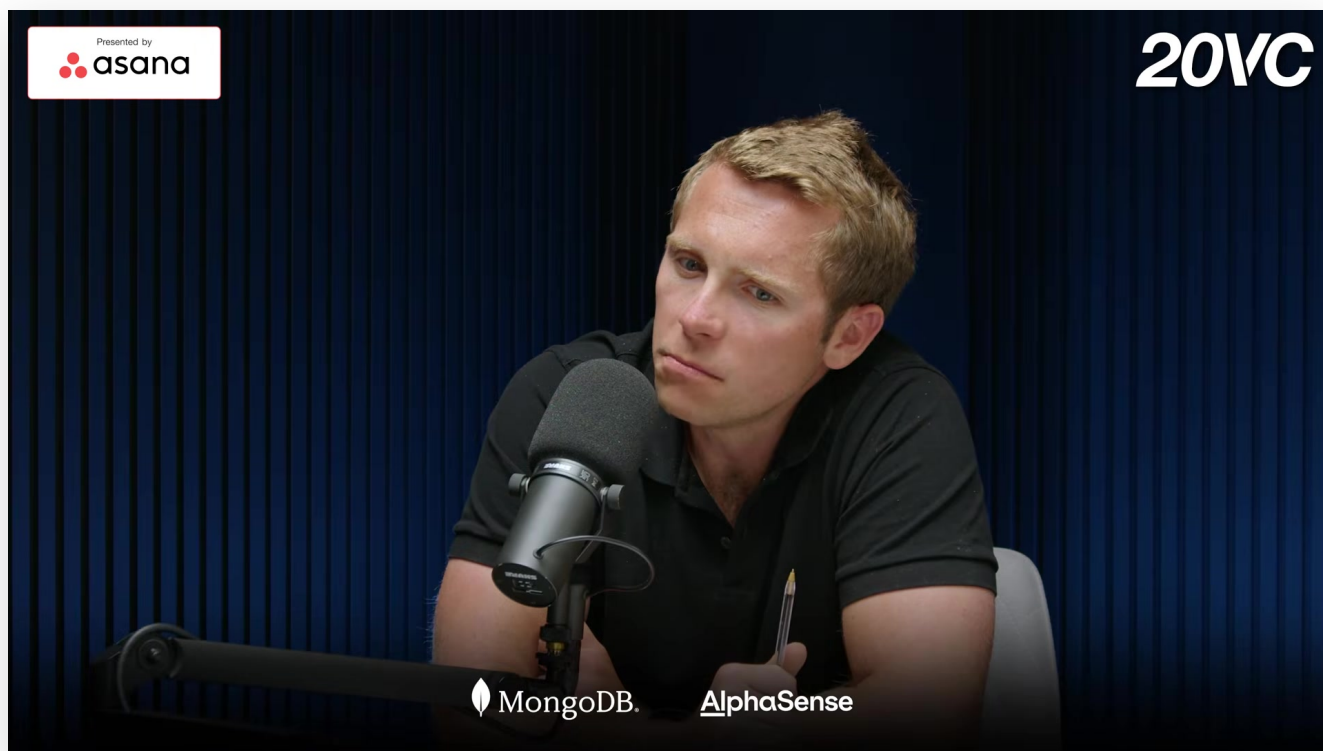
主持人随即问出第一题：对于今天学习计算机科学的人，你最大的建议是什么？

嘉宾开头回答说，学计算机科学没有问题，不要太焦虑，因为别人告诉你的很多事情……

画面说明

这一帧里嘉宾仍是近景发言，身体朝着麦克风前倾，像是在把争论从政治推演重新拉回到产业建设本身。这个时间段里，主持人先把自己对政治交易和监管打压的担忧推到极致，设想通过税、禁令等方式阻止中国开放模型扩散；嘉宾则把逻辑反转，认为正因为现在很多优秀开源模型来自中国，美国更需要自己做出开放模型，不能显得失去技术创新能力。他进一步给出支撑理由：美国开放模型会有很多顺风，湾区技术圈普遍知道这是个问题，而且包括英伟达在内的多方都在投入资源推动美国开源模型发展。主持人最后表态希望其成功，嘉宾则上升到原则层面，强调多模型世界对所有人都重要，然后话题自然转入快问快答。画面中嘉宾稳定地主导回答，恰好承接了这段从监管担忧转向“美国必须自己建设、而且多模型格局很重要”的论证。

37. 谷歌样本与创业资本错配 (54:49–56:21)



场景 37

完整内容

主持人问，在他看来，哪一家传统大公司采用 AI 采用得最好。Arvind 先反问一句：你愿意把 Google 算作一家传统公司吗？主持人回答说，愿意。Arvind 随后表示，如果算的话，Google 大概比任何其他公司都更靠前，不只是因为它在内部真正拥抱了 AI，也因为它不断把相关产品推向市场。不过他也补充说，把 Google 放进这个类别里多少有些不公平，因为它本身某种意义上就是一家 AI 公司。

接着主持人换了一个更个人化的问题：如果你重新创业，只能带走一个品牌大使，你会带谁？Arvind 回答说，他大概会从他们现有的大使里选一个，因为他们和所有人关系都很好。主持人追问到底会选谁，Arvind 则明确拒绝回答。他说自己确实没有现成答案，需要看具体情境，取决于自己要做什么，因为不同的人会带来不同的优势。

然后主持人又问，他最想改变当下创业生态中的什么。Arvind 说，他确实认为，如今创业公司可获得的资本太多了，而这有时反而在为人们制造失败路径。他觉得，这样的环境并没有让创业者真正得到建设一家伟大公司所需要的东西。随后他举了一个例子：一家刚拿到种子轮融资的创业公司，决定给

一位工程师支付 50 万美元年薪，就像他们前面提到的那样。这种事今天确实在发生，创业公司的创始人接受，投资人也接受，但这显然不是一条能够真正赢下去的可持续道路；他们愿意这样付钱，而 Google 并不是这么做的。

画面说明

这一帧给到主持人的近景：他穿着黑色上衣，身体前倾，一只手握着笔，贴近麦克风，表情明显是在认真追问。这个画面和这一段的结构一致，因为这里不是嘉宾单向长篇输出，而是主持人连续抛出几个短促但锋利的问题，把讨论从“哪家传统公司最会用 AI”一路推到“如果重开一家公司会带走谁”，最后落到“今天的创业生态哪里出了问题”。虽然主要观点由 Arvind 展开，但画面里的主持人状态准确传达了这是一轮高密度盘问。整段论证的核心也因此很清楚：Google 可能是传统公司中最成功拥抱 AI 的样本，但真正让 Arvind 更担忧的，是创业市场资本过剩，导致公司用不健康的方式烧钱抢人，例如给工程师开出 50 万美元，从而搭出注定难以持续的组织结构。

38. 资本过剩下的退出难题 (56:21-57:52)



场景 38

完整内容

Arvind 接着上一段的例子继续说，创业公司之所以愿意那样高价买人，而 Google 却不这么做，是因为 Google 知道自己根本不需要用这种方式去买人才。所以在他看来，资本过于充裕带来的一个直接后果，就是它把创业公司推向了一些注定无法持续的组织结构。

主持人随后问，他会不会担心如今退出路径越来越少，而且这个问题正在变得越来越真实。主持人解释说，他的意思是：如果今天一家公司的年收入不到 10 亿美元，坦白讲就很难上市；而大型科技收购方对自己想买什么也越来越具体、越来越挑剔；私募股权机构也还在舔舐伤口，因为它们的投资组合里装满了“烂摊子”。Arvind 回应说，这确实是一个艰难的环境。

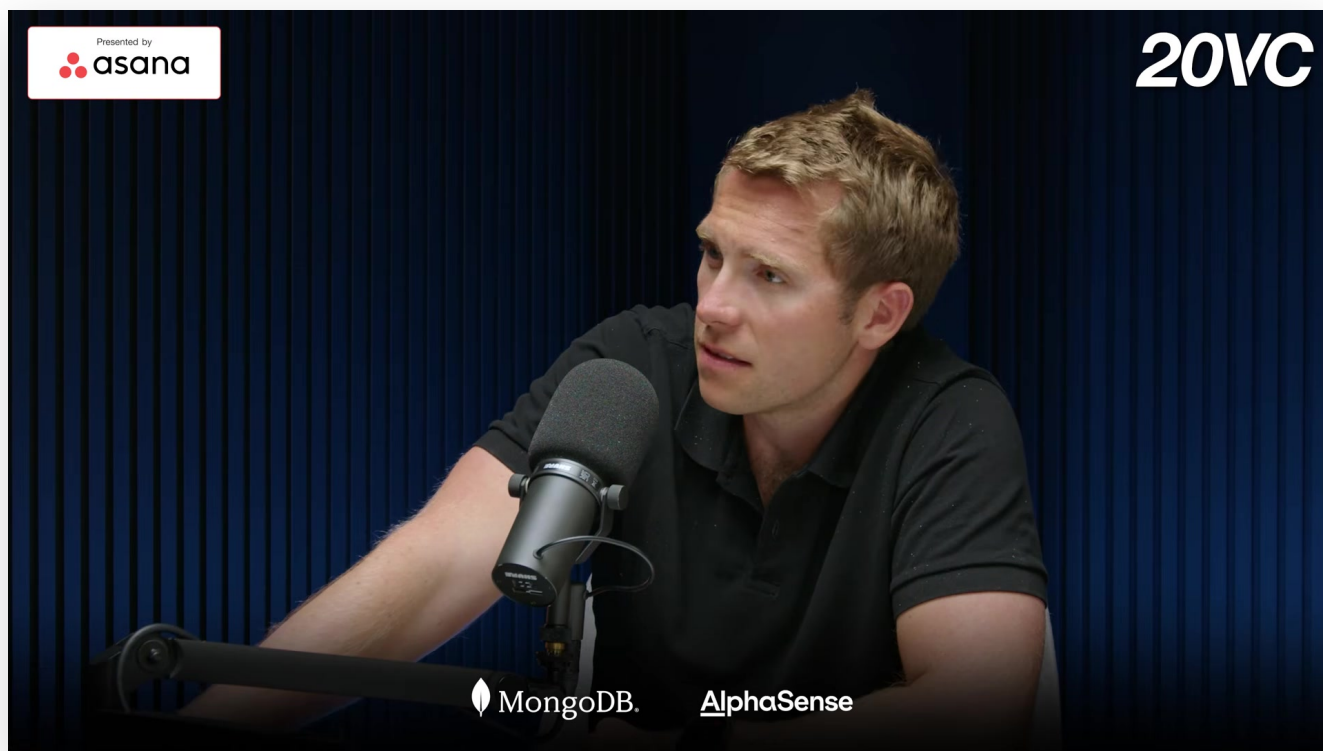
不过他也强调，创业从来就不容易。以他过去 25 年看到的情况来说，他反而会说，如今创办一家初创公司并最终获得一个不错退出，实际上比过去更容易一些。创业本来就是一场残酷的游戏。

接下来主持人又问，从外部看不到、但大家应该知道的创始人兼 CEO 真相是什么。Arvind 说，这根本不是一份性感的工作。事实上，这可能是压力最大的事情之一，你真的得有点“疯”才能做下去。

画面说明

这一帧切到 Arvind 的近景。他戴着浅色鸭舌帽，穿着浅蓝色 glean T 恤，双臂收在桌前，贴近麦克风，神情收紧，像是在给出不太讨喜但很直接的判断。这个视觉状态和他此处的发言是匹配的，因为他讲的不是乐观叙事，而是对创业生态的结构性批评。他先把前一段的薪酬例子补完整，指出资本过剩会诱导初创公司搭建不可持续的组织；接着面对主持人关于退出通道收窄的追问，他承认环境艰难，但同时拒绝把今天说成“史上最差”，反而认为过去 25 年里，如今创业并拿到好退出其实比过去更容易。最后他把讨论拉回创始人角色本身，强调 CEO 这份工作既不光鲜，也极度高压，只有足够偏执的人才能长期承受。

39. 创始人不该追求舒服 (57:52–59:24)



场景 39

完整内容

主持人接过话说，在他看来，还有一点是：你必须持续地不开心，你不该真正感到满足。作为 CEO，你永远都不应该觉得“已经够好了”，因为总有事情还需要被完成，也总有事情还能做得更好。

随后主持人又补了一句，说你告诉别人“你永远不会快乐”，这种说法常常会让人听得目瞪口呆。Arvind 回应说，这确实是个很好的说法。他接着展开解释：这份工作从头到尾都很艰难，而那些没有真正做过的人，往往会以为它充满光环，以为这会带来很多钱、让生活变得非常美好、让自己得到很多尊重。但在他看来，这里面几乎所有东西其实都无关紧要。要想作为创始人生存下来，你必须真正以使命为导向。

接着主持人问，钱会不会改变他的风格。主持人先给出自己的判断：如果说得直接一点，他觉得那些本来就已经富有的创始人和投资人，往往会表现得更好，因为他们做决策时不会带着经济上的急躁，会更理性、更稳健。Arvind 回答说，对他自己而言，也许确实有一点这个因素，但同时他也说，自己

是一个需求很少的人，而且这些需求很久以前就已经被满足了。所以，他一路做这些创业公司时，的确少了那种“我能不能养家糊口”的焦虑。也许这的确帮到了他；但随着他获得越来越多成功，这并没有从根本上改变他。

画面说明

这一帧又回到主持人的近景。他一只手向外伸，身体斜靠在桌边，像是在把自己的判断直接扔到台面上。这个画面很适合这一段，因为这里不只是 Arvind 单向回答，主持人也明确加入了自己的创始人观察，甚至提出了一个有争议的命题：CEO 应该长期处于不满足状态，而已经富有的人往往更能做出理性决策。围绕这两个判断，Arvind 一方面认同“创始人工作不光鲜、必须使命驱动”的核心，反驳外界对财富、名望和 glamour 的想象；另一方面也承认，自己因为个人需求低、较早没有家庭生计压力，可能确实少了一层经济焦虑，但他坚持认为，后续获得更多成功并没有改变自己的根本工作方式。

40. 创业者最后要靠不理性的执念 (00:59:24-01:00:01)



场景 40

完整内容

Arvind 把上一个回答收束完整。他说，从根本上讲，自己并没有被成功改变，因为你仍然必须具备那种驱动力，仍然必须持续工作。你必须比公司里的其他所有人都更努力地工作，要以身作则，不断往前推，而且你还必须带着一种近乎不理性的渴望，非得让某件大事发生不可。

随后主持人开始结束采访。他说，自己非常感谢 Arvind 愿意花时间来聊，也为自己在讨论中不断强力回顶表示歉意。Arvind 回答说，这场采访和主持人做过的很多采访都不一样。主持人接着说，这次对话更像一种来回展开的讨论，但他真的非常感谢这段时间，也觉得 Arvind 的表现非常棒。Arvind 最后简单地说了声谢谢。

画面说明

这一帧是收尾时的双人宽镜头。Arvind 和主持人分别坐在桌子两侧，彼此面对，中间是两支悬臂麦克风；主持人面前能看到纸张和杯子，整体空间安静、对话感很强。这个画面正好对应这一段的双重功能：前半段是 Arvind 对“钱没有从根本上改变我”的最后补充，他把创始人的要求压缩成几条最硬的标准，包括持续工作、比团队任何人都更拼、以身作则，以及带着近乎不理性的执念去推动大事发生；后半段则是访谈正式收束，主持人感谢并为自己强势追问道歉，Arvind 则认可这场访谈的不同之处。画面里的双人同框因此不仅说明说话权已经从单人陈述回到互相致意，也把整段谈话落在一种带有分歧但彼此尊重的结束氛围里。