

背景：日前，DeepSeek 创始人梁文锋在一场 4 小时闭门投资人交流会的深度分享 PDF 在业内流传。谈话实录中，他坦诚地讲述了公司的创立初衷、独特的愿景驱动式管理、对开源和商业化的看法、AGI 的技术路线图、以及面对竞争和资源限制的战略思考。本文未经梁文锋本人及 DeepSeek 方面确认。

以下为梁文锋四小时投资者交流会实录全文(完整版)

梁文锋

欢迎各位投资人。我们一开始来做这个公司，初衷是没有想到说我最后要赚多少钱，要到资本市场上去，要上市，要什么样的，所以我们是没这个初衷的。最开始的几十个人完全没有这么想过。如果他这么想，他就不会来。

所以总体讲，我们是怀着一个对这个世界非常大的善意来做这个事情，然后我们觉得这是对人类有用的，这是一个金钱以外的事情。当然，到了后面，这个事情发现利益非常大之后，又有其他的诱惑，这个是另外的事情。但是我们出发的初衷、我们的愿景，以及我们保持到现在的这个愿景，不是按照一个商业利益最大化的方式来做的。

我觉得这点比较关键。大概二十年前，在管理上我最崇拜的是杰克·韦尔奇，就是 GE 的前 CEO。现在回来看，他说的大部分东西可能都已经不对了，但是他最重要的一点说对了：一个公司最重要的是它的愿景。管理一个大公司，靠的不是你的规章制度，靠的是愿景。愿景是什么呢？

愿景不是挂在墙上的标语，愿景是你怎么做，不是怎么说，就是你怎么实际运行。反正杰克·韦尔奇说的原话我忘了，大概是这个意思。所以，我们是怎么管理这么多人、怎么组织起来的？其实我们没有组织，就是愿景驱动的，靠一个愿景来组织。我们是没有组织的。这个有好处，也有坏处。

未来我们会想办法扬长避短，但这是我们的特色。我们并不是以一个“我要实现什么 KPI、没有考核”的方式来做，只有愿景。这个愿景甚至也不是成文的，并不是写出来的，没有写出来过任何东西。这个愿景是在我们做事情的方法、我们对待这个世界的态度里。

可能我们公司每个人对这个愿景的理解也不一样，可能每一个人的愿景也是有差别的，但是在一个大的方向上是一致的。

我觉得还是怀着对这个世界非常大的善意，然后想做一点事情。我们是用这个来组织起来的。接下来我先讲，讲完之后大家再提问。我可能会围绕这个愿景来讲后面的事情。这个愿景是真的，不是编出来的，我们是真的这么想，真的这么做的。否则你没法解释我们的很多事情。为什么我们这么坚持开源？

因为这个愿景本身就要求开源。你没有这个愿景，你没法把人组织起来。比如说，**也开源，但是**的开源跟我们的开源不一样。**的开源有一种被迫的感觉，他们觉得这不是本意，但是对我们来讲，这就是我们的本意。然后在开源这个事情上面，我们一开始就想得非常清楚。

首先，第一个就是愿景；第二个，我们认为要把 AI 这个事情在商业上做成，开源是有好处的。这听起来有点矛盾，有点违反直觉，因为在历史上，开源跟商业化都是有冲突的。但我觉得 AI 跟以前不一样。

因为在历史上，一个软件公司，它的市场一年可能就几十亿美金，离开了，它就没有了，可能就只剩下几千万或几亿美金了。但是 AI 这个事情足够大，最终它可能得占掉人类社会 GDP 的百分之十，比如说。那么它其实是一个非常大的数字。

一个人独占这个事情，你不可能独占这个事情，你一定得跟别人分享，否则你肯定活不下来。这跟之前开源做一个软件可能是不一样的，因为那个软件的市场就没那么大。但是 AI 这个事情实在太大了。如果说我们想独占这个利益，那么一定是要被历史抛弃的。我觉得最主要这是一个客观的规律，这是一个历史观。

并不是说我不开源，我就能独占这个市场，这在理论上就不符合客观事实。你一定会遇到很多阻力，一定会有其他的方法阻止你实现这个目标。在这种情况下，我觉得不一定要按照传统的商业思维。你需要有一套机制来确保你自己能够获得的利益是有限的，那么你才有可能做成。需要克制，我觉得是需要克制。

如果说我们想在我们手上把 AI 这个事情做成，首先第一个，我觉得是需要克制的。不能够想着人类 GDP 的百分之多少都归我了，或者说中国 GDP 百分之多少都归我了。你越这么想，越做不成。

所以我们从一开始就觉得需要克制。你越克制，你越有可能能够做成这层事。这是一个商业上的考量，当然这是一个宏观的考虑。我觉得这是符合直觉的，至少是符合我的直觉，或者说至少我是真的这么想的。我们并没有非常多的其他优势，我们没有什么本事，我们并没有比别人有钱，也没有说我们人员比其他公司更好，其实没有的。

你想，我们两年前成立这个公司的时候，我们又没有很多钱，又没有很多卡，又没有什么知名度，又没有什么号召力，我们就是一群非常平凡的人。

我们真的就是一群平凡的人。如果说喜欢的一个叙事是一群平凡的人做出了不平凡的事情，而不是一群天才做出了不平凡的事情，这个跟我们的克制是很有关系的，跟我们的克制、跟我们的愿景是一脉相承的。那么，开源跟商业化会不会有冲突？我觉得在 AI 这个事情上面，你不克制，你就不起。

开源是属于克制的一部分，那么我们的克制不仅表现在开源上面，我们还表现在很多方面上。但总体上来讲，我们是不用考虑开源这个事，不用考虑克制这个事情。你越克制，可能就越容易做成，或者说至少到目前为止是印证的，到目前为止是能解释得通的。

否则没有办法能够解释为什么我们能够做成：我们并没有什么武器，起点又非常低，资源又非常少，我们的人其实也就是随机的一群平凡的人。我自己也就是一个大学毕业的学生，也不是最顶级的那个学校毕业的。这个克制也是我们愿景里的一部分。AI 这个事情太大了，利益太大了。

我们非常克制，只要能够做成，最后利益都会非常大。你随便分一点，利益就非常大，所以现在根本不用考虑拿这里的哪一部分利益、怎么拿，我觉得根本不用考虑这件事情，因为这个利益足够大了。你只要分一点点，就已经很足够了。所以我们之前说，我们只赚取一个合理的利润，只看你的意愿，而不是利润之大，这个是不一样的。

这不是我们的 API 定价。我们 API 定价觉得一个合理的利润，大概是我们到市场上买一批设备回来，十个月收回成本，我觉得这是一个合理的利润。

在当前的情况下，考虑到你有风险、还有前期的投入等等，如果一个服务器我们在财务上按照三年或者五年的摊销，但在商业上，我们觉得大概十个月收回成本，我们觉得够了，OK，我们觉得够了。所以这个是我们现在 API 定价的逻辑。我们的 V3.2Flash 跟其他的，都是十个月收回设备的成本。

这就是我们的标准。它其实不是利润最大化，如果利润最大化，应该把价格设得更高。因为在这个价格区间，用户的需求是没有弹性的，就是我价格再翻一半，或者说我价格再抬高一倍，token 的消耗量区别不大的。如果我价格再贵一倍，我的总收入接近一倍。稍等一下，我对一下。哎呀，太好了。

我跟你们讲个故事，就是我们的一款模型。一开始我们担心需求太多，所以一开始把价格定得比较高，团队里大家不是很高兴。后来我把价格又降下来了，降到四分之一，大家就很开心。我觉得这个才是我们真实的想法。

就是我前面说的愿景，我们还是让这个东西对人是有益的，而不是我们赚最多的钱，而是我们在能够赚到合理利润的情况下，大家都用得起。我觉得这次我们公司其他人的想法，就是我们那时候降价的时候，公司群里面很多人是欢呼的，大家都觉得很开心。因为这是我们花了这么多精力，这么用心把这个模型做好的目的。

目的就是能够非常便宜、效果非常好，能够让大家都能够充分地用。我们就觉得这很开心，这是我们的动力，这是我们的愿景，这是我们公司能够凝聚到一起去做这个事情的共识吧，这是我们公司内部的共识。这点应该比较特殊的，因为降价这样对于我们其他的竞争对手来讲，肯定不是个好事，他们一定不是欢呼的。

因为你的收入、你的 ARR，那个你降一半，ARR 就掉一半。对，这是一个我们不一样的地方。我们觉得这就够了。从我们公司内部来讲，我十个月收回成本，这个商业上我已经非常满意了。对于公司外部来讲，我们也觉得这个价格是大家比较开心、比较乐意看到的，大家是双赢的，公司跟社会、跟所有人都双赢的。

我觉得，OK，刚刚有人在屏幕上留言说，十个月回本这个利润太高了。确实是还有降价的空间，确实还有降价空间。在模型的优化上也还有空间，所以整体降价空间还是比较大的。但是这个成本，十个月回本，我们自己能做到，其他家做不到。像可能阿里或者腾讯，他不是我们的优化，他的成本应该是要比这个高好几倍的。

这里还是有很多优化的工作。刚刚说我们为什么不继续降价，是因为它没有弹性。就是我再降价，需求不会更多了，或者我再降价，需求增加得很少了。因为这个价格所有人都用得起了，大家都觉得这个价格是满意的，不会因为这个价格贵而不用了。

所以降价首先公司不会有更多的收入，对社会来讲也没有更多的价值，因为这个价格大家都满意了。你价格再低，对这个社会的幸福感也没有增加很多。对，OK，不过刚刚这个问题上，就是我们在定价这个事情上面，我们肯定不是以公司收入最高或者利润最高，不是最出发点的。这是我们克制的一部分，因为从短期来讲，你价格高一点，你可能收入多一点；但从长期来讲，还真不好说。因为我觉得克制是一种策略。对我来讲，克制是一种战略。就在于有时候你可以舍弃一些，来换更多其他的东西。

开源这个事，其实也是一样的，也可以认为是我们的压力，也可以认为是我们的让利。首先，这个让利对于我们公司内部来讲，我们很高兴，大家都很开心，员工觉得很有成就感，我们会因此有凝聚力。以及这个让利对社会是有好处的，社会也很高兴，其他同行或者说普通人都会很高兴。

所以这种克制，我理解是这种克制从长远上来讲，能够增加我们做成 AGI 的概率。在考虑一件事的时候，我是毫不怀疑 AGI 会有非常大的商业价值。那么在这个基础上，我优先考虑的不是我怎么多加一点份额，我怎么多拿一些份额，我优先考虑的是我怎么增加我能够做成的概率。这个克制可能还体现在很多其他的方面。

比如说，我们去年春节用户突然很多，但是我们并没有去追求我要留这些用户，或者说拿这些用户来变现，或者说我要去抢这些商业利益，在用户上面兑现。我们没有去抢用户，没有去赚钱，但我们很努力想办法把用户服务好。

我们并不会有这样的想法，说我要做成下一个超级 App，然后我要去跟谁竞争，我要做成下一个字节、做成下一个腾讯，完全没有这样的想法。我们可以这么做，但是我们没有这么做。我的理解是，这也是克制的一部分。你不要想什么都要赚了，你好像有了用户之后，好像就能够做成下一个字节了，然后你就把那个吃了。

我觉得这个在商业上是行得通的，是有可能的。如果去年时候我们用了大笔钱，就跟字节去抢用户，也是一种打法。但是我们选择是一种非常克制的做法，就是我不跟你去争这个东西，因为后面还有西瓜，前面的可能都是芝麻。

我不应该什么芝麻都抢了。当然，可能这个芝麻比较大，但我觉得后面的 AI 可能前面都不算大。现在来看，去年我们没有在 C 端去发力，可能是对的。因为能看到后面真的是有更大的西瓜，前面真的只是一些小芝麻。如果去年我就有很多钱，然后把这个事情做得非常大的话，有什么好处？你并没有得到什么东西。

这些是我的真实想法，因为我觉得后面的 AGI 机会应该是非常大的，后面的 AGI 机会永远是非常大的。

我甚至不用考虑我到时候在里面是占一个位置，或者说在那里面的商业模式是什么，我们根本就不用考虑。只要有那么大的商业机会，你一定是有办法的。那么前面的芝麻，我们也会捡，但是我们就随手捡一点，并不会说我停下来，或者说我把它当成重要事情来做。所以去年的 C 端日活这些，我觉得可能就是个小事。

但我们也捡了，我们也用一个比较低的成本维持了用户的使用，因为有可能以后是有用的。虽然现在不知道这用户有什么用，现在它是一个纯成本的支出，但是以后可能是有用的。既然是随手能够拿到的，我们就会顺手拿到。包括今年来看，很有可能我们在 API 或者 AI 方面的 ARR 收入，也是有一个机会的。

就是如果这个需求可以继续扩大，如果继续扩大，显卡可以买到更多的显卡，那么 ARR 做到几个亿美金是很有可能。如果说 AI 能做到十亿美金的话，那么基本上我公司的现金流可能就能够回正了，能够 cover 我的研发费用，能够 cover 我的所有费用。

所以说这个也是有可能的，但是我们没有把它当一个优先来做。这个我们会做，但我觉得这个是个重要事情。它不是我们第一优先考虑的，或者不是我们今天真的关心的。更大的机会应该还在后面，前面的机会包括去年的 C 端、今年的 B 端，我觉得这事要做，要把它做好，但这不是我们的目标。

或者说，我们公司大部分人认为这是一个非常重要的事情，不认为这是一个跟 AGI 比较起来同等重要的问题。开源那个可以多说一下，因为之前有很多问题问的都是开源。首先，我觉得我们是会开源的，然后我们最强的模型可能也是会开源的。因为我看不到闭源什么好处，看不到必然的好处。

字节它的模型是闭源的，它有什么好处？我看不到有什么好处。哪怕是模型开源，你把所有东西都告诉别人，这个门槛也非常高。别人要用起来，这个门槛也非常高。他要用起来，就很难；其次，他要用起来，还要成本做得很低，也很难，没有那么容易。

并不是我开源了，他就能轻易地做到跟我一样的部署成本。这里边还是有很多工作要做的。虽然说这些工作原理都明白，但是不是每一家公司都愿意，或者都有这个意愿和能力来组织人力去达到这个目标的。这点我也很习惯。它可能就不擅长做这个事情，因为它阻力太大了。它要控制这个成本很难，它有很多管理上的以及物理上的制约。

这也是创业公司的优势，因为创业公司如果太小的话，你没有这个力量来做这个事情；如果你是大公司的话，你很难组织。

这个事情都有难点，所以这是属于我们这个规模的一个甜区，**sweet point**。然后如果我们更大了，可能我们没有其他问题；如果更小的话，朋友们力量又会不足。所以，至于开源，我觉得我们应该定价。目前我们应该认识到，不会逼人。

因为定价的模型，我也不会收一个非常高的费用，我也是可能按照十个月回本来收这个费。按十个月回本，就已经能够把竞争对手.....我按十个月回本这个，就能够让独立部署的第三方无利可图。第三方他做不到，他做不到这个成本，他肯定做不到。所以开源并不会影响我的收入。当然，如果说我要赚一百倍的利润，那么开源是.....

主持人

听到你的话，但视频好像掉了，老板。

刚刚可能是电话接来了。

梁文锋

就是开源，我觉得对我们的商业模式是没有任何影响的。前提是我们只赚六倍的利润，十个月收回成本，大概对应的是六倍的利润。我们只赚六倍利润的情况下，开源是不会有影响的。但如果说你要赚一百倍利润，那么开源确实会影响你赚一百倍利润，因为第三方会部署，他可能是二十倍成本，就比你低了。

这个模式是不是长期可以持续的？我觉得是可以的。就在我们这个愿景下，我觉得开源是长期可以持续的，或者我们是打算这么做的。你可以认为，就是有克制，也是有让你比较长利。这个策略是能够在技术前面让我们有更多的机会，我们能够做成 **AGI** 的概率也是更大的。我们更从容。

你想，我们根本都不用加班，因为就没那么难。但是对其他家来讲，可能就很难，因为你想的也太多了。其实不难的，根本就不难。才开始是一个.....外面可能看起来，我们选了一个很难的模式，我们要做研究，我们去做最难的事情，好像是个 **hard** 的模式。

但其实我们在外地方舍弃了很多，使得我们还是非常有力的，我们还是做得非常轻松的。所以对开源这件事情，实际上我的判断是，它是持续的。开源和商业付费之间没有冲突，前提是六倍利润的情况下没有冲突的。六倍利润看起来很高，但其实不高。因为现在 **AI** 的效率这么高的情况下，现在一个合理利润可能就是这么多。

未来它可能会降到，比如说四倍、三倍，我觉得这已经是.....再到底也不可能再降了。但就是它还是会有很大的利润。就光是看卖 **API** 这个事情，但我并不觉得卖 **API** 这个事情有那么大吸引力。但是就这个事情，你可以说明，没有，我没有看到有什么冲突。

然后我也不担心别人部署我们的模型，然后跟我们来竞争，一点都不担心。我还希望他们能够部署起来。我们尽可能给开源社区提供帮助，协助大家能够把我们的模型部署起来。我不担心他会跟我抢这个生意，因为这个市场足够大。我只担心他部署不起来，他有些细节没做对，效果变得比较差，或者说他的成本会比较高。

对，这里是没有冲突的。然后去年的时候，我问的，去年有 **ToB** 这个生意的时候会问得比较多的是：我们那个 **C 端**，我开源，那么 **C 端** 是不是跟我这个 **C 端** 又会冲突了呢？

因为我没有流量的优势，或者说腾讯自己流量很多，他部署我们的开源模型，他就把所有的 C 端用户都接过去了，就把我的 C 端用户都抢走了。但其实并不会，这个还是有很多原因的。然后问题是：我们给的开源模型，跟我们自己部署的模型是不是一样？是一样的。

我们不会说开源一个差点的模型，然后我们自己部署的时候用一个更好的模型，是不会的，是一样的。这也说明它其实没有冲突。我们去年一整年，在 C 端我基本就是开源了，然后 C 端的服务没有看到冲突，真的没有看到冲突。所以，这是开源的这部分。

然后下面还有就是，公司的长期 vision，我觉得我们目标应该是 AGI。每个人对 AI 定义不一定一样，但是不妨碍我们把 AGI 当做我们的目标。从技术路线上来看，其实 AGI 这条路线图是比较清晰的。

跟目前的这一代 AI 技术，如果说你能够把一个问题描述得很清楚，给它完整的上下文和指令，它已经超过人类了。但这里有个定义，有个前提是：你给它完整的上下文，你给它完整的指令。然后这个定义是很难达到的。

比如说我们今天开会，其实我们前面有很长很强的上下文，可能大家有几十年的上下文，然后这是 AI 不具备的。那么现在 AI 能具备的是，在一个局限的上下文里面，它能做得比人更好。但它还是替代不了人类。这里面还差一个地方，是持续学习。因为人也能够持续学习。

你招一个员工，他可能花两个月时间来熟悉这公司的环境，熟悉他的工作，他谈两个月的，你这样他就可以上手了。他就能做很多事情。他能够听懂你说的话，比如说你说，叫那个小王来，他就知道小王是谁。但如果 AI 的话，因为这个上下文，他前面没有这两个月的学习。

你跟他说，叫小王过来，你得告诉他小王是谁、什么职务、他在哪里、要怎么去找他、要找他的时候注意什么。你得给 AI 所有的上下文。那么在这种情况下，AI 是能做的，但是你不可能给他所有的上下文，也不现实。所以 AI 并不能够替代你的员工。

但如果 AI 具有持续学习的能力，它跟你的员工一样，到公司学习两个月，那么但就可以替代天下的人了，所以我们离下一步还差一个学习去学习。AI 的发展，我们可以理解成它是一个阶梯。去年走的阶梯是 CoT，就是思维链。因为我们发现，通过思维链的方式，可以让智能达到一个更高的水平。通过它自己思考，可以让这个上限，可以让这个 AI 能做更多的事情。那么我们又跨过了一个阶梯。

今年的阶梯就是 Agent，因为我们发现，用 Agent 的方式，即便多有事情也可以做，它的能力范围会更大，它的智能上限会更高。为什么是阶梯呢？因为后面的每一步都是基于前面的基础上的。Agent 要用到 CoT，然后 CoT 也要用到前面的阶梯，前面的阶梯就是语言模型，所以它并没有一步是白走的。

所以，AI 的发展，智能的走向是有迹可循的。说今年走的这个阶梯是 Agent，但是 Agent 这个阶梯，它也是会走完的。就是它把所有可能解决的问题都解决完之后，但是它还是不能替代你的员工，但是它已经到它能力的上限了。

就好像 CoT 一样，CoT 到它的上限之后，它已经超过最顶尖的人类了，在做奥数题、写程序上面已经超过最顶尖的人类了。但是它还是停在那个地方，它那个技术并没有能够到达 AGI。所以你看，AI 这个智能的走向，它是有迹可循的。

然后在 Agent 之后，我们觉得应该要解决的问题是持续学习，就是怎么让模型可以持续地学习，而不是说你要给它一个很强的训练，它应该能够像人一样做一个比较长时间的持续学习。这个问题跟完成任务等等是同一回事，它们相关，解决的是同一个问题。

我们现在站在 Agent 这个地方，能看到的是下一个瓶颈，就是持续学习。下一个要解决的问题，就是解决怎么持续学习，这个是看得到的，是相对来讲比较明确的。就是卡在前面的这个障碍，你必须要跨过去，而且一定是有办法能跨过去的，但需要时间。持续学习之后，可能我们就会来到一个奇点。

这个奇点就是，当这个模型能够持续学习之后，它已经能够做人类能做的所有事情了。它就能够自己开发自己的版本，能够自己再研究，然后开发自己的下一个版本，开发更前的人工智能模型。所以它就会到一个奇点，能够实现自己的迭代。但这个奇点，它并不是一个奇点，它也是一个渐进的过程。

这个过程可能也是一个比较长的渐变，它不是个突变。但是习惯性地，我们都认为它可能是个奇点。因为在很早之前，那些预言家认为这里会有个奇点，但其实它不是一个奇点，它是一个连续的过程。然后这一步走完之后，我觉得才是具身智能。

这是我们的推测，这是我们觉得这个时间表应该是：先解决学习去学习，然后再到那个智能的奇点，能自我迭代的奇点，然后才是具身智能。到具身智能之后，它就走进现实世界，可以给你做家务，可以给你养老。我们觉得这是一个比较理想的路线图，但每个人的观点不一样，没有对错之分。只是我们觉得，这个路线图是最轻松的。

这个路线图是因为每一步你要做新的都很少。这个路线图我们可以不用加班。但是如果路线图是反过来的，比如说它要先实现具身智能，那么这里自己做得很累，这是一个很苦的活。我们不希望是这样的路线图，我们希望做得轻松一点。

如果说我们先解决持续学习，再解决那个自我迭代的奇点，再解决具身智能，这个路上就很轻松。因为到后面之后，你可以用前面的技术来帮助开发后面的技术。奇点之后再做具身智能，那个就不用人来做，就不用我们来做，那个模型自己就可以出来了。所以这个是回答我们的长期目标是什么。

我就跟他说，这就是我们的长期目标，就是我们说的 AGI。大家回到现实。去年最重要的现实就是，大家都要做 Chatbot，要抢 C 端的流量。那么今年的现实是，大家都要抢 To B 的收入，要参与到这里面去，因为如果不参与的话，根本就不在牌桌上，对不对？

但我们并不认为它是一个重要的事情，或者说，在我们公司内部，我们真正关心的，其实是刚刚我说的这些 AGI 的路线图，以及下一步这个技术怎么突破。但是有一点很奇怪的是，最想得到的东西，反而你就得不到。那个并没有那么在意的事情，反而是还蛮容易能够得到。

这里边有一个战略上的优势，就是我们心里面想着 AGI，我们做的是 AGI。那我们再做那个应用，再做那个 C 端、B 端的时候，根本就不用太多的心思去做那个事情，其实花很少的精力就可以了。我觉得是说，你站在一个技术的高位上来做相对低一级别的技术，是有这样降维打击的。

至少在去年的 C 端，我们看到确实是这样。我们没有在 C 端上花很多力气，甚至一度我们都不想维护那些用户了，但是用户赶都赶不走。因为真的是赶不走，所以最终都还在。但稍微.....然后今年 B 端的这个收入，现在看起来这个增长还比较乐观。我觉得可能这个数字跟同行比的话，应该也是比较好的，我估计。

但是我们并没有花很大的精力去做这件事情，我们根本就没有做一件事情，就是顺便做的。做互联网智能的上线，走的时候，AGI 这一步是我必须走的，是个台阶。通往 AGI 的路上，我要经过这个台阶。那么我把这些技术都通过 API 给大家服务，我没有干额外的事情。我们还是在做 AI，这是一个副产品。

我只要搞几个人维护这个 API 就可以了，甚至客服都没有，也不用销售，什么都不需要，用户自己就会来。

或者说，被认为是 C 端的用户，C 端和 B 端，都是我们做 AGI 路上的副产物，都是一个中间的产出，跟我做 AGI 没有冲突。我并不是为了做 C 端，或者为了做 B 端而去做它，而是我们做 AGI 是为了做 AGI，刚好能产出这个东西，我就把它拿来商业化了。这跟其他公司不一样。

其他公司就是为了做这个，为了服务 C 端用户，或者服务 B 端用户而去做这个模型。但对我们来讲，初衷不是这样的，我们初衷还是去追求 AGI。我觉得这个在一定程度上是一种降维打击。AGI 是更大的愿景，这个愿景能够凝聚起更多优秀的人，它有更强的凝聚力。

所以我在组织上有优势，然后我利用这个优势去.....它就是一种降维打击。但如果你是一个商业公司，你的愿景就是服务好 C 端用户，那么是另外一个故事。他有其他的优势，在产品上、用户服务上、流量上会有优势，但是在技术上没有优势。现在比较有利的形势是，模型技术是最重要的。

你要将模型做好，其他.....对，然后这个可以解释我们之前发展的轨迹。我们真的是选择 AGI，然后我根本没有想着说要做很多的用户。去年春节我们突然火的时候，那不在我们的剧本里面，我们完全没有想过那个东西。我们就是想把这个技术做好。

我觉得我们之前秉持这种态度，其实我们并没有因此而少拿到任何东西，并没有因为我开源，并没有因为我们的善意或者说我对其他人提供帮助，而导致我们少拿了任何东西。比如说去年的 C 端用户，我们现在 C 端用户还是比较多的，还是比较稳固的。我们今年的 B 端，我觉得也是比较乐观的。

我并没有因为我们的善意而影响到我的商业利益，完全没有影响，反而可能还有加分。这个看起来违反直觉，但它确实是这样的。或者我们反过来想，如果违背它，也是必然的，我能够拿到更多东西吗？没有。有个问题是，怎么理解世界模型和 AI 智能上限提高没有关系？我说的是在现在这个阶段，这是我们的判断。

我们看到，这是我们自己的 AI 路线图，不是唯一的路线图。从我们的理解、从我们的判断来看，当前最重要的是把 AI 训练做好。把 AI 训练做好，并不需要世界模型，甚至不用多模态。因为你把 AI 训练范围缩小一点，没有多模态，只是一部分任务你做不了，但是不影响这个算法的成立。多模态最终还是要做的。

现在重要的是训练，然后下一步要解决的是持续学习的问题，再下一步要解决的是它自己问问题。但是这个路线图里面没有世界模型，没有视频生成。视频生成一开始出来的时候就很火，好像这是必须要做的，如果你不做，你就不是一个 AI 公司一样。

所以我就很奇怪，这个其实你只要仔细去想一下，它跟智能的路线图是没有什么关系的。

事实上，你也发现，那个视频生成一开始 Sora 出来之后，所有人都做，大公司、小公司都做。但是小公司后来都把它砍掉了。它跟智能的上限没有关系。但在商业上，它是个好生意，在商业上是个好生意。但是这跟智能没有什么关系。我们不会因为它是好商业而去做它，我们只会因为它是智能路线图上的东西，才会去做。

视频生成那个，因为那个相对比较清楚，就拿它举例。然后世界模型，世界模型这个含义就没那么清楚，因为很多东西都可以说它是世界模型。从我们的判断，世界模型和智能还不是现在这个阶段最重要的事情。最重要的才是 AI 训练，以及 AI 训练之后怎么解决持续学习。

这是我们公司的判断，当然每个公司它的判断是不一样的。我刚说的是，我们公司对公司来讲最重要的问题，就是人员的稳定性。从另外一个维度来讲，我们缺什么呢？我们跟美国的差距是什么？其实差距只有一点，就是资源。我们没有那么多卡，我们的卡的数量还是比较少的。

我们现在大概是两万张 H 等效力，这其中大部分是刚到，最近一两个月刚到，可能还有很多机器还没有来。我们总的算力去年比较少，今年我们在非常激进地扩充这个算力。我们现在大概是两万张 H 等效，接下来几个月我们还会有大批量的机器买过来，基本上都是英伟达。

我们需要多少卡？现在肯定是越多越好。在我们能够承受的范围内，肯定是卡越多越好，这是毫无疑问的。所以我们现在策略是，在合理的价格里面，能买到多少卡就买多少卡。如果说这笔融资用完之后，我能买多少卡，我就买多少卡。花钱的速度不在计划里面，而是只要价格合理，有多少我都买。

那么如果说我在半年之内就把钱花完了，我觉得是个好事。如果说我在半年之内把钱都花完，那这个太幸福了，这太理想了。实际上，要把这么多钱花完非常不容易，买不到那么多卡，很难买，而且价格也很高，也不能说花非常高的价格去买，还得确保这个价格是合理的。

如果我半年之内就把这个钱花完的话，可能是最理想的。因为我把钱变成英伟达卡，肯定比放在银行里面好。放银行里面，我现在已经可能是存两个点，好像就要。但是买英伟达卡，十个月的社会成本。

所以肯定是你能够买到多少就买到多少。最先如果买了卡之后，后面我有很多空间，我通过提供服务或者我怎么样，我总是能够有现金流的。我有现金流我就可以活，我就不需要在我账上面放很多很多钱。所以说，我们只担心买不到那么多卡。

如果能够把钱都变成卡的话，那我们会毫不犹豫把所有的钱都变成卡，并且在这里面我们是愿意付一定的溢价的。就是我们愿意付一定的溢价去把它变成卡，因为这太划算了。哪怕我们在付完溢价之后，其实我们也很难实现这个目标。那么客观上来讲，如果说我今年能够花掉两百亿，那么就属于我们的采购部门业绩超级好了。

我们跟美国的差距主要在资源上面，然后人上面差距不是很大的。人上面几乎没有差距，因为就是同一批人，可能是中国人。中国人出去的时候，有一些人留在国内，有些人留在国外，有些人去国外，他并没有说是聪明的人去国外，没有的。其实它还是比较随机的。

最聪明的那些人，可能不是说一半多一点去国外的，但有一半少一点留在国内。国内人才是不缺的，而且我们的基数大，我们每年有那么多人的补充。人才不是瓶颈，资源是最大的瓶颈。资源首先影响到人才培养，因为算力少，我们能做的实验机会比较少，所以我们的人才整体上比美国有差距。人才的差距，本质上也是因为算力的差距。

在现在最大的模型上，我们其实叫不起的。我们哪怕把五百亿全花掉，其实也叫不起。哪怕能堆起来也用不起。现在最大的模型，它激活大概是八百 B；国内的话，我们还在几十 B 的这个规模，国内最大模型可能就几十 B 激活，那么差一个数量级。

如果我要训练跟 AI 同样大的模型，应该需要五万张 GB300，或者华为 950，二十万卡。这只是训练，还没有考虑做研究。所以我们跟美国之间最大的差距是在资源上面。我们现在的资源，以及我们今年之内、接下来几个月的资源，包括马上大资源，也只不过我们在十 B 激活的这个规模里面去做更多的实验。

因为我在几十 B 激活的这个规模里面，还有很多实验要做，还有很多事情需要搞清楚的。我们应该离能够训八百 B 的模型还比较远，时间还非常多，没有那么多卡。所以我们跟美国的区别，我认为就是资源上的区别。

我们可能会认为，我们看到的所有区别，包括人才的区别、模型能力的区别、应用的区别，可以认为都是因为算力资源上的区别。

算力资源一方面是国内本身卡就买不到，另外一方面是我们的资本投入比美国要少。我们在资本投入层面上少了很多，基于人才的工资在这里面占的比重是很低的。你看他们开的薪水一个亿美金这种，但算起来，人才的薪水还是占比很少，大头还是算力。这个问题目前基本上无解，因为华为的产量也是有限的。

因为我要训八百 B 的话，我就得二十万张华为最新的卡，这只是训练，还没有考虑做研究。所以我们现在根本就不会考虑，到这么大的一个规模上去跟美国竞争。我们现在输，还是在我们能够训练得起、能够用得起的规模上，就几十 B 激活的规模上要先把它做好。

再等下一步有更多的资源的时候，再把它做到一百五十 B、一百五十六 B，或者两百五十五激活的这个规模。所以我们跟美国在这里面有一个目前看起来很难弥补的差距。你要硬要训那么大模型也是能训，但是你没法做充分的研究。就是你在训之前，没法做充分的研究。

还有个大家比较关心的问题是，所有大模型竞争，它终局的差距会体现在什么地方？就是当大模型最终差距会体现在哪里？我觉得这个差距最终可能是没有太大差距的。最终的差距应该是三方面：一方面成本，一方面时间，一方面用户体验。除此以外，可能是没有什么差距的。

成本比较好理解，你提供同样的服务、同样质量的服务，你能够以什么成本来提供。同样一件服务，比如说比亚迪的电池，在同等技术量的情况下，其他家是不是能够按照这个价格来提供，我觉得这个是个较难的事情。不是那么容易做到的，肯定是壁垒。所以成本肯定是一个差异，我觉得可能成本是排在第一位的区别。

然后第二个就是时间，你什么时候能够做到。你早几个月、晚几个月，它就不一样了。第三个可能是体验，用户体验还是有些不一样的。这个用户中间还是会有一些用户粘性和用户壁垒的，但它可能不本质。本质还是第一个成本，第二个时间。你先做出来还是后做出来，你做到同样东西，是快点还是慢一点。

长期商业化路径和产品线进一步丰富后，怎么定价？我们觉得现在最值得做、最值得花精力的，还是做 AGI。现在要把 AGI 往前再往前推，就是把智能的下限往上推，往前推。这个应该是在现阶段，比做更多的产品线、考虑更多商业化路径更划算，或者说它是一个收益更大的路径。

我觉得在未来的一段时间，以及过去的任何一段时间，可能都是这样的。就是说，如果说我们花了很多时间去思考，丰富我们的产品有什么用吗？半年前讨论出来的商业化是什么？一定是我要去做广告，然后我要去做电商，我要去在产品里面植入电商，然后要跟本地生活什么很大一体。肯定没用的，因为变化太快了。

你这个产品，如果说你在前面，或者我们现在这个阶段，去花很多时间做商业化考虑、商业化路径，产品线它生命周期很短。我觉得没有到这个时候。这是我们的判断，或者至少前面的这些经验都是支持这个判断的。在过去三年，任何一个时候你来跟我说商业化路径、产品线，都是浪费时间，因为你并不能够预测、不能够预见未来。

你能够预见的是很少的。因为我看时间，不知道大家要中场休息一下吗？要先吃点东西吗？还是我们继续聊？大家没有意见，我就继续说。在做全球领先的 AGI 研究，并在适当时候进行商业化。我觉得我们一直在做商业化，我认为一直在做商业化，只是没有以商业化为目标。

我们是以 AGI 为目标，但是我们一直在做商业化，所以我们才有 C 端的用户，才有 B 端的收入。从历史经验来看，这个策略是成功的。然后我觉得，我们离彻底转向商业化的时间点，应该是很遥远的。所以最大的还是技术的延伸，然后做下一代的技术，更多地解决现在的问题。

这些收益比，我个人觉得，就在现在能看到的未来，在任何一个时候你聚焦在产品上都太早了。所以这也是我们克制的一部分。我希望这些商业机会能够让其他人来做，我们希望这些商业机会、怎么用这个 AI，是整个社会、我们所有的合作伙伴一起来做，大家一起来分享这个收益，而不是我想独吞，这不可能。

而且我们没有那么多精力，我们的组织也没有那么多人去做这个事情。对于合作伙伴，其实我们这个融资是精心挑选的。具体怎么合作的这个提议，但首先我觉得，利益是比较一致的，就是跟我们利益最一致的、对我们最没有敌意的，或者说最希望我们能够做成功的。

不是所有人都希望我们能做成功的，因为我们还是挺帮了很多其他人利益的。公司重大战略、技术业务决定的流程和决策机制。我们公司整体上是建立在共识的基础上的，我并不是说我一个人决定所有事情，而是我要寻求共识。我在公司内的权威以及在公司内的影响力，是建立在共识的基础上的。

比如说我要做一个事情，我肯定是先看我们大家的共识是什么，大家想不想做。然后有可能我会有一些引导或者倾向，但是这引导的作用是有限的，引导的作用是非常有限的，还是建立在一个共识的基础上。

这个决策机制其实是一种寻求共识的机制，这并不是说我能够推动一个什么事情，它一定是共识，我才能够推得下去，然后我才会去推。目前主要精力基本上都在 DeepSeek 这边。

主持人

我们休息五分钟。

投资人

我快点写。

主持人

大家可以开麦，给我一点反馈。

投资人

我们这边没问题，要不然先休息五分钟。

梁文锋

各家模型最终效果拉开差距，应该是一个综合上的。比较模型效果，肯定是得在相同的成本上来比较，这个才是有意义的。因为你比较两辆车，也是同价位的车来比较。做得好的模型跟做得差的模型，这个差别应该不是在具体某个环节，它应该是整体上的。Anthropic 跟现在超过 OpenAI，这是不是长期的？

我觉得这不是长期的，这肯定是极端性的。OpenAI 和 Google，未来大概率还是会交替上升，应该是会交替上升。其实现在 Anthropic 在 Code Agent 上的优势没有那么大，其实并没有说它碾压 OpenAI。

我们公司可能有一半的人，平时有一半的人觉得 OpenAI 是更好的。其实 Anthropic 它有先发优势，但这先发优势应该很快就没了，并不是一个它能够长期占得住的优势。大家这三家都很厉害，这三家里面的效率是最高的，它花的成本、它花掉的、它烧掉的钱应该是最少的。

在全球 AI 中主导分工的时候，中国公司很有可能扮演的一个角色还是产量最大。常理来讲，我们的产能最大，包括芯片，芯片可能我们的产能最大，我们的电力最多，所以我们的 AI 最终很有可能我们是三体之一。

中国人会把这产品做到最便宜，然后再在效果上，毕竟国外的商品，现在很多商品中国产跟美国差并没有太大的区别。未来可能 AI 也是这样，但是中国产的 AI 可能价格会更便宜。这个便宜可能是系统性的低，就跟其他行业中国提供的服务更便宜可能是一样的。

我要做事情的时候，我习惯的思考是：我现在这个时候做什么收益最大？如果说我觉得现在做产品收益最大，我觉得去做产品；如果说我觉得现在把 AGI 先实现收益最大，那我就是先去做 AGI。显然，我觉得现在做产品不是收益最大的。如果你是国产卡适配的问题，国产卡现在其实是有一个历史性的机遇的。

因为之前国产卡适配有一个难题，叫生态不好。就买了卡之后，但是用不起来，它没有英伟达那个生态。所以说英伟达的护城河是很强的。

但是这个事情在发生变化。英伟达 CUDA 的护城河在快速地被瓦解，快速被瓦解的原因可能是有三方面。一方面是现在有了 AI，然后有了 AI 之后，我要建立起这个生态比以前容易很多了，因为 AI 可以写代码。

我可以用 AI 来构建这个生态，就可以把跟英伟达一模一样的生态构建出来。第一个，因为有 AI；第二个，有一些新的技术。比如说，我们家出了一个技术，叫 TileLang，是一种高级语言。

用这个高级语言来写 CUDA 的算子，可以很快地把英伟达的整套生态全部都写一遍，再结合 AI，这个看起来没有什么障碍。但是现在还没有完成，还没有做完，不过这个技术路线看起来是没有什么障碍的。

还有一点，因为 CUDA，英伟达它是从游戏卡演变出来的，所以它在很多地方，游戏卡的设计跟游戏卡的设置是一脉相承的。CUDA 是兼容游戏卡的。以前因为 AI 计算是一个很小的领域，比游戏卡的市场要小，所以这样是合理的。但是现在计算卡的市场已经比游戏卡更大了，就没有理由这两个还需要耦合起来。

现在的趋势是，以后就不再耦合了。那么专用芯片，不管是华为还是英伟达自己，以后都是专用芯片，都不是之前的这些东西了。在这个背景下，原来英伟达生态的作用就大幅度减少了。因为对于一个专用芯片来讲，跟 CUDA 没有什么关系，它跟 CUDA 是不绑定的。

或者说，这个芯片在设计的时候，就已经考虑到怎么建立这个生态了。有点复杂，但 anyway，国产 AI 芯片替代现在是有个历史性机会的。我们认为，未来一年之内，我们能够看到一个事情被验证：国产芯片的生态完全没有问题。

之前认为是有问题的，认为是用不起来、不好用，但是未来我觉得一年之内，我们能够扭转这个认知，或者会用事实来扭转这些。国产 AI 芯片的硬件和生态都没有问题，唯一有问题的是产能不够。国产卡适配这一点，没有障碍，英伟达挡不住。

如果是在一个正常的商业环境里面，我能买到英伟达的卡，那么国产替代是比较难的；但是在英伟达的卡买不到的情况下，所有人都迫不得已，都要去做国产芯片。在这个背景下，国产卡的适配是没有任何障碍的。国产卡建立起跟英伟达一样、甚至比英伟达还好的生态，我觉得没有障碍，但还需要时间。

我们现在主要是跟华为有合作。华为他们自己适配，但我们自己会参与这个生态，会深入参与到华为这个里面去。华为的问题还是产能不足。像华为给我们的大概是一万六千张卡的产能，互联网大厂可能是十几万张，我们一万多张，我觉得这个比例也是比较.....但这已经可能是华为就这么多产能了。

所以我们也没法指望在华为上面训后面那个更大的模型，或者说训练几百 B 参数激活的模型，有时就有说在今年。但是明年、后年说不定是有机遇的。华为卡适配，我们主要做的工作是把它的高级语言编译器做好，把 TileLang 做好。把 TileLang 做好之后，问题可能就迎刃而解了。

这个事情说起来比较复杂，但是我们在做，做完之后再解释会清晰比较多。你可以理解，V3 训练的时候，它用的还是英伟达的卡，但是已经不用英伟达的生态了。

V3 用英伟达的卡，但是没有用英伟达的生态，而是我们先写一个高级编译器叫 TileLang，然后基于 TileLang 的生态来完成其他所有的事情，就已经几乎不依赖英伟达的生态了。只要我把这一套东西，把这个过程重新在华为卡上做一遍，那么就完成了。

我觉得这里可能是一个历史性的使命，即可以完全扭转之前大家对国产卡生态不好的认知。现在有些天时地利基本上都全了，就差时间。我觉得一年之内，应该会有很多人在上面都有，或者说都明白这问题算是解决了，剩下就是产能的问题。我对国产算力是比较乐观的。我觉得在这一点上，英伟达是在掘自己的坟墓。

华为的超节点，华为的 950 超节点，在性能和价格上可以完全平替英伟达的 GB200、GB300。价格肯定要贵，但贵得有限。价格贵百分之五十、百分之一百，贵百分之一百无所谓，贵百分之两百都无所谓。比如贵百分之一百，我觉得已经可以认为在价格上可以平替了。

在任务上也是可以平替的，所有 GB300 能做的任务，华为超节点都能做，延时什么都一样。唯一的代价是，四张华为卡顶一张英伟达的卡，同时落后两年。四张顶一张这个可以理解。落后两年的意思是，四张华为 950 能顶一张 GB300。落后两年是时间上落后两年。

华为 950 超节点是今年 Q3 还是 Q4 的货，英伟达 GB200 是两年前 Q3 的货，差两年。英伟达今年 Q3 可能已经有新一代了。所以我们跟美国在芯片上的差距，我认为生态上以后不会再有差距，但是在芯片上是四倍加两年。

还有问题是，我们会不会向上游进行垂直整合？我希望不要。所以有个问题是，我们是不是会往上游应用进行垂直整合？我们希望不用这个，我希望是其他人来做这个事情。我不希望我把所有东西都吃了，我只要吃一块就好，我只要吃我最擅长那一块，或者我觉得我们自己认为最核心那一块，然后跟用户直接相关那一块，我觉得可能对于我们的很多产业伙伴来讲，他们比我们更关注类似.....应该是别人的意义，不应该我把它诠释了。未来我们会不会自建大型的集群？我觉得自建大型的集群是肯定要的，我们自己一直在做这个事情，我们所有的集群都是自己建的。

但是未来要不要自研芯片，我觉得取决于这里的收益有多大，取决于收益有多大。Tesla、培育.....这个事情。假如说你是运营发电厂的，你并不一定需要去造发电机，对不对？发电设备可以是别人造的，只要它的价格合理，你为什么要求自己造？所以，我希望不用去做芯片。我希望能够以合理的价格买到芯片，这样我就不去做芯片。

我觉得这个很有可能是这样的，就是最近英伟达芯片的利润，不见得是可以.....虽然.....我们希望只做一块。我觉得 AI 这个事情很大，并不需要我.....我只做一块。如果聚焦，并且我认为这里的生意利益已经足够大，就如果是 AI 时代会产生很多家万亿级别的公司，我觉得我们是其中一家。

我们一直做其中一小块，我们是其中一家已经没有什么问题了，并不需要我.....我并不是信心勃勃要做地方。我觉得最可能是比较难受，且你真的想做到的话，你还是会有更多的主意，这个会.....这是我们的态度。

譬如说，至少在 To B、To C 这两个业务上，目前能看到的，真的想做 To C 闭环的，反而我们做得好；真的想做 To B 闭环的，说不定也没有我们做得好。你想要得越多.....然后还有，To B 这个可以多说几句。

To B 业务的上限应该还是需求，在现在这一代 AGI、AI 技术的背景下，To B 的需求应该是有限的。它会快速增长，但并不是一个无穷大的事情，最终还是受制于需求，不是算力。在当前技术的条件下，收入有多大，最终还是取决于需求。

因为你这样想，我十个月就能收回成本，我肯定是如果有去处，我一块买，是因为没有那么多事情。对，需求应该会越来越大，如果随着技术持续有突破，这个需求会越来越大。

多模态布局，我们一直在做。对产品来讲，它很重要；对 C 端用户产品来讲，它很重要。但是对智能的上限，它是一个组件，它不是主线本身。但是它作为一个组件，多模态我们肯定会做，而且我们也在做。我们应该会上相关的模型，就是我们 V4、V4 的后续版本会支持原生的多模态。

但是我们对多模态、对智能来讲，它是个组件，我们不把它当作智能本身，它的主线跟搜索一样。搜索也是一个组件，多模态也可以，我们的理解它也是一个组件。Scaling，我们是信 Scaling，肯定是规模越大，效果越好，能够解锁更多的功能。

阻止我们 Scaling 的其实就是算力，并不是我们不想 Scaling，是我们没有那么多的算力去做这个 Scaling。我们没有触摸到这个上限在哪里。我们训练这么大的模型，并不是因为我觉得这么大的模型就够了，而是我刚好有这么多资源。

我是按照我的资源来算，我能够接受、能够训练的场景是多大，是这样算出来的，并不是这个模型就够了。目前来看，模型收益还是非常明显的。我们还没有机会碰到这个 Scaling 的墙，这离我们还很远。

硅谷在说 Scaling 到头的时候，那是对硅谷来说：对中国人来讲，我们离那个还很远，我们根本就没有 Scaling 到那个程度。这个 Scaling 包括数据的 Scaling、模型规模的 Scaling，然后训练成本。我们离探索这个 Scaling 的上限还比较远，就是没有那么多算力。

但我们也会不遗余力地去推进这个 Scaling 的上限。

包括我们融资完之后，我们有更多的算力，可能训练更大的模型。时间内买到，就在越短时间内买到；如果能半年精力全花完，那么这是最好的，实际上做不到。下一代模型的核心能力，我觉得它必须要有持续学习的能力，它才能叫下一代模型。在那之前，我们能做的就是成本，然后效果做得更好，速度做得更快。

但是要有大突破，它应该是具备持续学习的。然后还有一个问题，就是为什么好像我们特别在乎这个模型的计算效率？因为我确实发现，不是所有人都很在乎这个模型的效率。一个商业化公司来讲，没有动力去追求模型的效率，因为模型效率高一点.....所以，它没有那么大的动力去追求模型效率要非常高。

所以几个创业公司，他们自己并不会.....你没有听到他们说低成本是他们追求的东西，因为那个不符合他们的利益。低成本了，你还赚什么钱？低成本了，你就收不到什么钱了。

相反，这是我们愿景的一部分。或者说，我们的愿景，我们的同学在乎这个成本，因为我们的同学都是普通人，他知道用这个都是要花钱的。他能够有这个同理心：别人要花钱来用，那么别人如果便宜一点，别人能够接受的程度会更高。所以我们有很多同学还是希望我们能够把成本做到更低。

但是如果你从商业角度讲的话，其实不会这么考虑。从商业来讲...

从成本或者从商业角度讲，这不是第一优先级。对于不管是创业公司来讲，还是大公司来讲，这个服务成本根本就不高。但是我觉得我们希望它是比较轻的，我希望它是一个负担得起的，特别是在中国算力紧缺的这样的背景下，是负担得起的，能在国产卡上用的。我觉得低成本首先是一个结果。

我们的模型确实一直在模型架构上往一个更低成本的方向走，这跟我们的愿景有关系。我们还有很多在算法上的方法，成本还可以往下走。成本往下走还有一个原因是，成本越低，我就越能训练更大的模型，我就越能承担起更大的模型。在同样算力上，在算力有限的情况下，如果我的计算效率更高，我就能够承担起更大的模型。

对大公司来讲，他不一定会这么考虑。对大公司来讲，资源是可以加的，可以通过加资源来解决。但是我们会优先考虑成本效率。数据类型型的价值，数据这个范围比较广，数据应该几乎就等于模型的一半。

为什么我会觉得，如果我想要那个，或者假如设 AI 能够占 GDP 的百分之二十，然后如果说我想要在这里面占百分之五，绝对行不通。因为我肯定会被另外一个人打败，如果另外一个人说我只要占百分之一，那么肯定就会被他打败。如果说我的目标是，我要占 AI、全人类 GDP 的百分之五，理论上算这个账还是成立的。

你看 OpenAI，他算这个账好像是能算得过来的，理论上是没问题的。但是他有个问题，他会被另外一个愿意只占百分之一的人打败。因为另外一个人说，我做得这么好，但是我只要拿全球 GDP 的百分之一就可以了，那么就会把他打败。这时候如果又出来另外一个人，说我只要百分之零点一就可以了，那么又会把前面的人给打败。

从宏观上来讲，不管这百分之几是从哪里拿，彼此之间没有区别，都是一样。拿得多的人会被拿得少的人打败。甚至你不用真的拿得多，愿景如果是拿得多的话，你就会被愿景是拿得少的人给打败。其实大家都没有拿到钱，然后只是一个愿景。你愿景是拿得多，你就先输了，你就会面临着更大的困难。这个世界就是这样。

OpenAI 从一开始觉得他真的能够垄断这个世界，但是实际上他会遇到很多很多挑战者。他会遇到挑战，他就不会那么轻松。美国会遇到过的挑战，那么他在未来可能还会遇到中国的挑战，因为中国人愿意拿得更少，就可以给你提供这个服务。在中国，也会有人愿意拿得更少一点。

但最后这里会有个平衡，因为你拿得太少之后，公司商业逻辑就不成立，就活不下去了。所以你拿得太少了，你活不下去；你拿太多了，你会被拿得少的人打败。所以对我们来讲，我们并不是利润要拿最多的钱，或者说算收益最大化的定价，而是只赚一个合理的收益。这是一个解释。

我是相信这个事的，我并不是去为这个事情找理由，因为没必要找理由。我本来就这么做的，我这么做肯定是有理由的。这个理由可能不是非常惯常，但是我觉得公司本来就不惯常。我们公司的管理其实是两条线：一条线是从上到下，一条线是从下到上。从下而上，就是每个人自己想做什么，自己做，没有人管他，没有 KPI。

从上到下，就是做正事，我们要集体做一个什么事情，需要全公司的人一起来配合。比如说我们要发 V4，那么就得分工，每个人得做一部分。那个是从上到下，然后从上到下这个我们叫做“做正事”。一般我们希望“正事”不要占用员工所有时间的一半，就“正事”不要超过一半。

他还有一半的时间，是不被安排的，他想做什么就做什么。这是一个研究的范围，让他可以自己去探索，按照他觉得什么重要，他去探索什么，没有前置的要求。只要公司能够支持，公司算力能够支持他做，或者说他不需要算力，他需要做得很少，那么他根本就不用来协调。所以我们现在是这样的一个组织方式。

有些人觉得我们是从上而下的，有些人觉得我们是从下而上的，我觉得两个都对。我的一个标准是，“正事”最好不要超过一半。

我们一般也不太加班。加班有两个原因。第一个是，做研究是需要一个比较松弛的环境。你如果逼得很紧，就没法做研究。因为既然就是要你自己有这个兴趣，你自己平时要去想这些问题，所以得是在一个比较松弛的环境里，才有可能能够探索。这是一个出于研究文化的需要。第二个是，我们非常聚焦。

我们非常聚焦，就意味着我们要做的事情很少。那我就没那么多事情要做，我就不需要加班。这个跟前面的克制是一脉相承的。因为我克制，所以很多时候我就不做了。那么我要做的事情少了，每个人分到的工作就少了。你看到我们的产品很多都不完善的，我们也没有去补它。这也是我们的一种文化。

OK，因为问题很多，我大部分都扫了一遍。大家还有什么问题，大家提问。

主持人

请各位投资人自由开支交流吧。我稍微提醒一下，梁文锋讲了很多一些比较敏感的信息，请大家千万不要外传一些数字或一些情况，包括卡量什么的。同时也千万不要录屏对外做分享。很感谢各位，请各位有问题的话，自由开支交流。

投资人

杨哥，您能分享更多关于持续学习什么时候带来突破的时间线吗？包括如果实现持续学习，还需要哪些架构、算法的创新，以及其他关键要素？

梁文锋

人少，需要研究。现在全世界都在研究这个问题。或者说，对投资人来讲，现在投资人看到的最多的是 AGENT；但对于我们这些研究的人来讲，现在看到更多的是学习，以及怎么解决学习这个问题。其实学习可能不是一项技术，它是一个问题。怎么解决这个问题，可能有很多种技术，不是一项技术，它不是一个东西，会是很多东西。

或者说，AGI 是由很多东西组成的，需要模型，然后还需要很多其他东西。其实它也是一个工程和算法问题。这个问题比较专业，但有很多方法，也有很多研究。

投资人

梁总，谢谢。非常感谢今天这个机会。我首先非常想回应和感激一下，您最开始说的让我非常感动，也给我们很多启发。您提到，这个团队带着最大的善意，希望能够在这个行业里面，对这个社会、人类智能的发展有所推动，做一点点贡献。

并且在这个里面，是带着这种使命感和愿景。我觉得这跟我们所服务公司的企业文化非常相近，就是“修己达人”。我深刻地理解了为什么您会带领团队做开源的事情。我会形象地感觉，像我们去做一个榕树这样的小鸟天堂生态，利万物而不争，但这样它就会被大家接纳，万物共生共存，最终就会无处不在。

所以，我们会以这次投资，同样表达我们对这个使命和愿景的认可、支持和尊重。同时，我们也希望能够在产业未来，我们所擅长的一些领域等，去贡献一些力量。在这块也想跟您继续请教和探讨。比如在未来生态的共建当中，现在开源之后，这个行业里有多少伙伴、人才、团队，能够比较好地复现咱们开源目前的一些模型和成果？

未来在下一步想让这个生态进一步发展的时候，您感觉在哪几个方面，需要更多高质量的人才能够衔接到我们的模型，把它复现？还是说现在 GPU 的算力相对有一些稀缺？大家未来会不会是一种模型矩阵的方式？比如咱们把大模型基模做得越来越好，各行各业的伙伴和团队去做一些模型矩阵里面的垂直行业模型，或者一些应用模型。

这块目前的发展怎么样？未来一步一步，两年、三年，您感觉会长成一个什么样的生态？这是我第一个想跟您请教的。第二个，您刚才也跟很多伙伴分享了很多关于 AI 硬件方面的观察。像 AI 全球大公司，可能单体都会做千亿美元级别的投入，中国目前看起来在硬件算力上有些短板。

您感觉这个多长时间可以解决，并支撑咱们 AI 的发展，让算力和硬件短板不给 AGI 拖后腿？您觉得这是不是中国人未来使命必达，我们肯定能做出来，只是时间和资金投入的问题？但同时，可能它会是两方面的。

一方面，模型的进步会让模型智能化的提升，导致单一任务或者某些智能化单体对硬件、对算力的消耗逐渐递减，不再需要那么大算力的运算，因为模型的进步会让它巧算。我不知道我理解得对不对。另外一方面，硬件的技术进步会让算力的算能效更强大。这会不会是两边相向而行的路径？

现在如果是在这个时点，用现在的 960 也好，还是 H200 也好，去做千亿美元级别的算力投入，您刚才提到说您是给它按三年摊销，那它的实际生命周期，您觉得技术迭代是四五年？或者直白地说，会不会现在算力中心按现在的卡去建了万卡集群，可能三年之后它其实就是相对不那么先进的算力了？

会不会有现阶段是 under construction、不够用，三年之后变成相对不那么优质的算力有冗余的情况？我不知道会不会有这样一种现象。以上两个问题请教您，谢谢。谢谢。

梁文锋

第一个问题是生态的问题。我们现在觉得，可能每个企业都面临的问题是人才不够。但我觉得这个人才短缺会是阶段性的。我们在每个行业发展的初期，人才都是不够的。包括以前做网站，刚开始做网站的时候，做网站的人很少，人才很缺。后来互联网要做服务端，人才也是很缺的。

但是这种人才短缺都非常快会被解决，也就两三年，因为会培养出大量的人。AI 人才的短缺也是阶段性的，并且我们已经看到，大幅度被缓解了。因为 AI 人真的不缺，每个公司很快会把人培养出来，培养人是很快的。所以，AI 这个行业整体上，不管是生态、模型公司还是什么，人才都不缺。人才缺肯定是一个短期现象。

历史上从来没有出现过长期缺一类人的情况。我还记得十几年前说飞行员很缺，飞行员的培养周期也很长，但也很快被解决了。所以大家不用担心缺人才的问题。以及国内现在做模型的公司有点太多了，还是太多了。美国可能就三家，中国做基模的东西太多了。最终一定是不需要那么多人去做基模的，一定会收敛。

所以资源也是比较分散，某种程度上也比较浪费。就先于每一家都要做同样的事情，但美国只要三家做，资源只集中在这三家。中国资源分得很散，每一家拿到的资源就更少。我觉得这个肯定是会收敛的，一定会，但这需要过程，最终一定会收敛。不需要那么多家，因为现在可能大家觉得做这个事情的利润率非常高，所以一定要自己做。

但当他发现这个事情可能没有那么高利润的时候，可能就不做了。最近肯定是没有那么高利润的，我不相信有那么高利润，因为这不符合客观规律。这意味着我们是处在一个阶段上：如果有一个非常高的利润率，这一定不符合客观规律。我们应该是一个合理的利润。所以这是产业的一个现状，我觉得肯定会收敛。

就是大家做大模型的那一部分，其中不要说某一家独占，说“我要拿走全部利润”，这个肯定不行。如果说每一家都只拿合理的利润，那么其实不需要那么多人去做大模型。中国最后有个三四家竞争，竞争就很充分了，价格绝对已经够打价格战了。

大模型可能不说两家大公司、两家小公司，可能就已经比较够了。至于生态上的，我没有什么太多的想法。我们希望能够扶持更多的人，但是我们并没有那么多的精力。我们是有这个意愿，并且不会有利益冲突，但是我们有没有去做是另外一回事。但至少这里还是没有利益冲突的，我们是希望合作共赢的。

首先，我绝对不认为大模型公司可以拿走大部分利润，这个不可能，因为这么多家大模型公司，现在差距不用那么大。差距只有两个东西：一个是时间，一个是成本。所以不至于哪一家有暴利，我觉得不至于有暴利。成本控制得好的人就多赚一点，成本控制得差的人就少赚一点，仅仅此而已。是不是有回答了？第一个问题是不是回答完了？

投资人

大家能...你相信以后肯定是有很多人可以...未来其实会...大家数据应用这些的循环迭代...现在可以听到吗？谢谢。对，感谢您的回答，也非常深刻地理解和尊重您的这种行业里面的生态战略定位。比如说数据这一块，现在公开的数据，相信模型公司都已经可以有渠道获取，这个方法应该都不成问题，只是时间和成本的问题。

那么后续比如说将到真正到 AGI 的时候，有可能大家一个设想或者理想的状态是，模型可以自我迭代、自我学习，就是自己训练自己。那么这一块的话，目前这个数据，您感觉仿真数据是不是可以用起来，还是说真实数据的质量最高？

如果说还是需要来自于真实数据，那会不会是限制这个 AI 的智能还是在人类的过往...这个层面，因为它依赖的是人类真正曾经有过的真实数据？还是说可以突破这个上限，通过模拟数据、仿真数据、创造数据等等的方式，让这个模型能力去超越人类过往的所有真实...

梁文锋

梁文锋我觉得是能超越的。我觉得有两点超越的，比如说围棋，AlphaGo 他下了一手人类从来没有见到过的棋。就是说，他肯定是在一定的范围内超越人类的。但是他可能也有上限，他可能也是有局限性。但是这个局限性我们现在看不到。

我们认为，所以笼统地认为，它是可以基于人类已经有的、我们已经能说出来的知识上，予以超越的。

投资人

投资人那这块后续是靠真实数据还是仿真数据？它会 work 吗？

梁文锋

梁文锋不可能就有很多方法。

投资人

好的，谢谢。也占用您的时间了，也想继续请教刚才关于 Allinfra 的问题。

梁文锋

梁文锋第二个问题是什么？有点...

投资人

好了，我简单快速地重复一下。就是想请教，对于 Allinfra 这一块，未来相信算力现在大家都是千亿美元级别地在投入。那么这块的话，有可能我们相信中国人未来在硬件上是使命必达，有一天可能会有高效率的算力，但有可能这个在实践的过程当中，目前还是一个肘。那么未来会不会是两面向下而行？

一方面是模型的能力迭代之后，它其实对于算力从硬算变成巧算，所以单位模型或者任务对算力的要求和消耗会逐步地边际降低。另一方面，比如说硬件像卡的能力提升，会让迭代速度越来越快，单卡效率提升。那么这个在过程当中是怎么样一个现象？

会不会是说，现在去建了万卡集群，去买了 H200 或者 960，但是过个两三年，它就变成了一个相对没有那么优质的算力，相对又变成了一个陈旧的器件？

梁文锋

梁文锋英伟达的卡基本上你可以按照五年折旧。华为的卡最多按三年折旧吧。华为 950 今年能用还挺好的，明年用我觉得还可以，后面再用我觉得可能就真的太费电了。华为卡生命周期肯定是会短一点，因为它本来就已经比英伟达晚两年了。但是我觉得差距没那么大。如果说 B200 现在能买到多少，我觉得都划算。

假如说对于腾讯来讲，阿里巴巴能买到的话，再看量；如果合理价格能买到，肯定都是划算的。但不是个算成本的时候，相信买不到。

投资人

明白。

梁文锋

我们算力落后，这是一个事实。这个事实通过三方面来消解。第一方面是我们承受模型落后，我们只能够用比他们更小的模型，小多少问题就是训练。我们要承受一定的模型落后，以及模型的尺寸更小。这落后有一个好处，落后意味着你有更多的时间，你有一定的技术，然后这样的话你就可以用巧妙的方法

投资人

谢谢。

梁文锋

所以我们跟美国的差距可能是落后美国 12 个月，落后美国可能 12 到 18 个月，或者说 6 到 12 个月。反正简单说，就是落后美国两年，然后只用美国二十分之一的算力把这个事情做出来。这个叙事就是落后一到两年，但是只用它二十分之一的算力。

那么未来我们要把这个叙事改写，就是我们用它几分之一的算力，但是把这个时间缩得更短，缩到 6 个月、3 个月，我觉得这是一个目标。以及我们甚至可以在某一些方面超越他们。但是在总体算力还是有数量级差距的情况下，全面超越是不现实的；但是在某一些重点、有取舍的地方，我们有一些地方超越可能是可以的。

投资人

明白，谢谢梁总。信心满满，一起努力。时间也留给其他的伙伴，谢谢。感谢刚才的分享。我这边有两个快速的技术问题。在刚才聊技术路线当中，提到了我们这一阶段要解决的核心问题是持续学习，也是目前国外研究的热点，叫 RecursivelyImprovement。想请教一下，目前来看，从技术上最大的难点是什么？

从您的视角来看，这个什么时候可以解决？这是第一个问题。第二个问题，同时您刚才提到，先解决持续学习，然后再去做智能。我也想了解一下，您这么提背后的技术根源是什么？是不是意味着解决完持续学习问题以后，DeepSeek 后续也会做通用智能？这两个问题请教一下。

梁文锋

技术问题其实解释起来有点.....它的难点在于，我们现在还没有找到非常 work 的方法。全世界都还没有找到好的方法，大家都在摸索。所以现在还在摸索阶段，就是不知道谁能摸到下一个解决这个问题方法。现在还在探索阶段。我们有很多思路，有很多现在看起来有前途的一些想法，但是都还没有做通。对，这是第一个。

第二个是，现在我们内部比较看重这样一个叙事：训练我们的下一版模型，我们希望它能够帮助我们自己的开发。它能够提升 DeepSeek 的效率，我们的模型最首先是提高 DeepSeek 自己的工作效率，让我们开发下一版模型的时候，它能够提供更多的帮助。

或者说简单一点，我们做的模型，第一目标不是大家用得好用，而是我们自己用得好用。首先是对我们自己有用。对我们自己有用之后，我在开发下一版模型的时候就会更快。

我们内部很多人的想法是这样的：首先要对我们自己有用，首先是给我们自己用。然后这是实现 AGI 最快的方法。当我们自己好用，那意味着可能别人也好用，但是首先得保证我们自己好用。这个叙事有点奇怪，但是确实很多人就是这么想的。

而不是说用户用得好用，我是希望对我们自己帮助更大，这样我们可以实现 AGI，会快很多。所以这个叙事的逻辑是，来帮助我们实现 AGI。但首先是帮助我们实现。我们实现 AGI 需要这个帮助。现在是非常确定，我们确实需要人工智能来帮助我们实现 AGI。

虽然它还不是自主地工作，它还是只是跟人搭配，但是已经很有用了。

投资人

第二个问题，就是刚才提到了，先解决持续学习的问题，然后再进入通用智能。这是您对后续的一个预期吗？想理解一下这背后的技术根源是什么？为什么先要解决持续学习，然后再进入通用智能？您对这块后续的理解。

梁文锋

因为解决持续学习这个问题，可以大大加快我们的研发进度。如果我先解决了持续学习这个问题，那么通用智能这个问题就不在话下了。我有了 AI 的辅助，如果 AI 能够持续学习，它的能力应该是非常强的。现在的 Agent 的能力受限，是因为它不能持续学习，它不能有效地持续学习。

如果说能够先把持续学习做完，那 AI 的能力是非常强的，它能够非常大地提升我们自己研究的效率。持续学习先做出来，通用智能可能就很容易了，用它来做就很容易。所以我说这是一个我们比较希望看到的结果，我们比较省力，我们就轻松。

否则现在你要去人工做通用智能，它是一个比较累、比较苦，是一个数据密集、人力密集的事情，性价比也不高。

投资人

感谢分享。

主持人

小问题，线上的问题你看一下聊天群。觉得 AGI 还需要多久？国内硬件在这个时间能追上吗？在 Zoom 会议的那个聊天窗口里面。

梁文锋

好，这个我看了。华为 950，现在华为是给我们一万六千卡，这应该是可以公开说的。应该是比互联网大厂少一个数量级。华为也只能给我们这么多，因为这个价格也不便宜。

互联网大厂的追求会更大，对互联网大厂来讲，它更需要。对我们来讲，我们可以买一些不合规的卡。所以我们买华为 950 的目的，还是希望帮华为把这个生态做好。一万六千卡的华为 950，只相当于四千卡的 B 系列。所以说不是一个很大的量，意义不是很大。

不够训一个下一代的模型，它只够训我们现在这一代模型，不够训下一代模型。但是可以让华为把这个先做好，就是关于华为 950。然后，AGI 还需要多久？国内硬件在这个时间能追上吗？

我觉得在 AI 这个事业上面，在 AI 这个事情上，应该国内一两年是能够做到跟国外差不多的，或者可能今年就能做到平替国外的模型。在 AI 这个事情上，就现在的一个方式、现在这个范式，不是很难的事情，所以今年应该就能做到的。但它还不是 AGI。

我至少觉得，它得能够持续学习吧。国产硬件在这个时间能追上吗？我觉得国产方面可能需要一个几年的时间。国内首先得解决生态的问题，因为生态是一个信心的问题。解决生态的问题，然后再解决产能的问题，我觉得应该是能够逐步解决的。我不太相信未来五年之后，我们还卡在产能的问题上。

现在肯定是卡在产能问题上，今年、明年、后年，我觉得可能都还是卡在产能问题上，但五年之后，我觉得可能不一定，我还是比较乐观的。然后第二个问题是，思考未来的组织架构，对人员的规划规模。首先，我们之前的组织架构是非常分散的，因为没有组织架构。但是我们人在进一步扩大之后，这些肯定是需要做出一些改变的。

现在只能说，这里会需要做很多改变，但现在很难一下子全部表达出来。最后应该还是要分不同的部门，应该有些部门是我们需要建立起比较严谨的层级结构的。另外一些部门，可能还是会维持一个比较松散、比较扁平的结构。随着人员的增加，我们会做这个调整。应该是我们马上就做得这个调整，因为我已经在做这个调整。

已经不做个调整的话，很多事情是没法推进下去的。确实有很多部门，是应该有组织架构的。然后大家还有问题是，CV 是先于它的哪个版本是吧？我觉得我们现在，线上发了这个 GCV4 这个版本，还是比较粗糙的，然后能力方面还很多都需要时间。

我们一般，对我来讲，一个比较舒适的发版节奏，大概是两三个月发一版。上次发版可能是四月底，那么下次发版可能是六月底，大概是这样子。如果没有意外的话，应该还是会一版比一版好的。在 50B 激活的这个尺度上面，我感觉最终我们跟现在的这波开源不会有太大差别。

就在推理速度跟效果上，我感觉可能就不会有太大的差别。但是跟它那个大尺寸的模型，就它一个没公开的模型，应该差距还是比较大的。那个差距应该是，我们这个激活参数应该是做不到的，这就肯定得是一个更大尺寸的模型，可能比如说是 150B。

那么 150B，我觉得按照我们现在训练的进度，今年，乐观的是今年年底可以开始去训练，至少是明年回....差距还是比较大的。对，这是跟 OCE 的差距。

投资人

你好，我其实有一个问题，就是经常您说到这个 AGI，它实现的过程是一个渐进式的过程，而不是会有一个突变。所以我可以理解，这是一个没有临界点的过程吗？

梁文锋

它没有临界点，但是它是非线性的。我们现在比较相信一个叙事是，AI 可以加速 AI 的研究，AI 可以加速 AI 的研究。就是说，它不是线性的，因为你可以用 AI 来加速你自己的研究，所以它到后面可能是非线性的。

投资人

明白。所以当您的这个，我理解前面的结论可能就是，语言模型继续 scaling 是够的，是足够的，做到这个状态。

梁文锋

只能说语言模型的 scaling，我现在没有看到有上限。我们现在的智力水平，或造成美国的智力水平，都没有看到上限。

投资人

明白。因为我非常好奇，就是你说，对美国来讲，800B 激活这个模型，说比它能训练起，但它也用不起来，所以它只能训练出来，它也很难拿出来给大家用了，因为确实有点贵。我之前其实很好奇的一点就是，人类其实拥有语言能力就是近十万年的事情，然后前面可能进化了三十七亿年。

但在训练 AI 这个事情上，可能是可以，你说那个顺序是可以反过来的，但最终可能还是会进入到所谓，也不一定叫世界的模型，还是要进入到物理模型或者具身的那部分，是吧？就是在这个上限之后。

梁文锋

对，我觉得具身肯定还是要进入，最终具身。所以对我们公司来讲，自然而言，可能终点都是具身。因为对一个正常人来讲，他的需求并不是电脑，对不对？因为正常的人，他吃喝玩乐、衣食住行，他不需要电脑。

他需要的是，所以他还是需要具身智能来解决具体人力的需求。如果说目标是解约人力需求的话，所以具身我觉得就是绕不过去的。

投资人

明白。所以阶段性地讲，假如说到达类似，也不一定是临界点，就是可以自进化、可以比较好的自进化，或者接近这个与 SV 的这个上升的到那个时间点上，我很好奇会希望这个状态的 AI 落地的第一个让是.....可能跟现在不一样。

梁文锋

我们是希望它能够，就是直接假如说没有具身，那么我们对 AGI 的定义，或者我们希望 AGI 能做什么呢？它能帮我迭代下一版模型，能帮我迭代下一版模型一样。然后如果有了具身之后，我们希望它做的也是，让它来迭代下一版的具身，它来做下一版机器人。

投资人

我还是很好奇一点，就是因为我看之前 DeepSeek 的采访等等，应该是在一些重要的方向的选择和研究的选择上面，品味和直觉是很重要的，而不是简单的工程优化。那如果之后 AI 可以自进化的话，这个品味、taste、直觉这些东西还重要，或者说重要的会是什么？

梁文锋

AI 现在不缺品味和直觉，它缺的是持续学习的能力。AI 的品味和直觉没有问题。你让它写篇文章，它的品味和直觉，我觉得没有什么问题。前面还有几个问题，我看一下那个问题。我看屏幕上谈了几个问题，但是我这边看不到了。

投资人

梁总，我屏幕上留了个问题，我给您再念一下。其实是想请教一下，continuous learning，就是持续学习，您也提到，很多研究员也提到是一个还没解决的研究问题。然后那个 coding agent，尤其是追上 MILES，就是到这个 Office 水平、MIS 水平是一个比较确定的目标。对于一个还没解决的研究 Scaling 目标和一个比较确定的 Scaling 目标，您认为应该怎么分配研究资源，尤其是研究员这块的人才资源，怎么样才能达到最好的平衡和效果？

梁文锋

模型 Office 是一个比较确定的目标，但是模型 MIS，我觉得还不好说是确定的，只能说是一个目标。模型 Office，我觉得应该还是比较确定的。

CoT 不占资源。做朋友任意研究，他不消耗卡，只需要很少的卡，他需要的是想法；他也不消耗人才资源，因为不需要有人一直在那里做。他不是一个项目，而是需要有很多人都在那里想这个问题。所以并不需要给它分配什么资源，因为它不需要资源。你要调模型，要做模型、发模型，做模型的效率实验才需要资源。

刚才讲的，对人、对卡的资源消耗都不多。所以我们叫这个叫“摸奖”。门槛很低，谁都可以去摸，但是谁能摸出什么来，这个可能我也不知道是看天赋还是看什么。所以这里并不需要我们去分配资源。只是说，我们跟其他公司不一样的地方，就是我们会花时间去讨论这个问题，会去想这个问题，然后把它当做一个重要的事情。

在公司里，它是一个重要的问题，是一个我们会花时间去思考的问题，但是并不需要花很多资源去做。然后下面我还看一个问题：大模型的幻觉问题比较影响用户的体验。幻觉问题也是有一个方法可以解决的，但是这是一个长命题。

幻觉问题可以认为是一个可以通过更好的 Post-training 解决的，是一个能解、能够改善的问题。

只是大家没有花很大的力气去做。或者说，对我来讲，幻觉是一个问题，但我们归结可能是产品问题。我们会去解决它，但是不是重点的问题。前面还有一个标数据的问题。我们在数据标注方面，这跟我们的资本投入有关。以我们这个资本投入的结构，支撑不起那么多高质量数据标注的成本，因为成本很高。

美国数据标注的成本跟中国数据标注成本没有什么区别。中国去标数据并没有成本优势，尤其是标高端数据上并不会成本优势，使得我们很难投入去像美国这样标数据。这条路在中国是很难的，因为标数据实在太贵了，不管是我们外标还是我们自己标，都很难受。所以现在基本上是两条腿走路。

并不是说我们完全不能标，而是因为标数据有一些成本低，有一些成本高。我们先标成本低的。所以你也可以认为，现在在我们公司有一半的人在标数据。有一半的核心研究员，最重要的人，有一半在标数据。我们就集中在标数据。解决 AI 这个问题，在现在这个阶段靠的就是标数据。你只要看就是一个数据。

投资人

梁总，谢谢你。你讲得特别地，我们非常有感受，非常好。感谢你。我请教几个问题。

第一个问题，您刚才讲，中国的模型比美国的模型在效率上肯定是强的。你讲的还有一些其他方面，未来也有可能会比美国强。你觉得是哪些方面，我们在智能或者在其他方面会比美国强？

梁文锋

梁文锋我觉得在很多体验方面，有可能我们是能比美国做得好的。即不说自己的体验，自己的产品体验觉得挺好的，我觉得我们在用户体验方面不一定会比美国差。在产品方面，产品能力上不一定会比美国差。成本应该也会比美国低，所以有可能中国还是会有竞争力的。其他方面，如果说有结构性的优势，我觉得可能也没有。

但是在成本和产品这两方面，我觉得确实是有一定的结构性优势的。成本这个很好理解，是因为他们都不用做，所以他们就不发展这个能力。他们肯定没有我们重视这个事情。我们可以把它当做一个非常重要的事情，但对于他们来讲，这个是不重要的。产品也是，原生通过很多公司，产品能力还是可以的。

所以我觉得这两方面可能是有结构性优势的。

投资人

好的。第二个问题，请教您，您刚刚讲到后训练，我们投入相对成本又高，像 Anthropic 和 OpenAI 都投入巨大的金额。这次融资以后，您觉得我们会在后训练方面加大投入吗？

梁文锋

差距主要是在高质量数据的标注上，然后主要是在 AI 研究里面。肯定会加大投入，但是像高质量数据的标注，典型不是资本投入。高质量数据标注的瓶颈，我觉得是时间，就是需要时间。因为对 OpenAI 来讲、对国外来讲、对 Anthropic 来讲，他们都更早，然后资本更多，卡也更多。

这种情况下，我们国内可以认为是最近半年才开始做，所以在时间上，我觉得是需要更多时间的。这个跟资本投入关系不太大，因为哪怕没有更多的资本投入，原来的资本也够它以最快的速度在扩张了。但是这个速度是有上限的，瓶颈不是说我能够马上有更多的人，并不卡在钱上，也不卡在卡上。但是它确实是在一个快速扩张的过程。

所以我们觉得一年之内，高质量数据这个问题做得比较好，我觉得国内应该是可以预期的。我觉得面可能没有那么冷，但它确实是需要时间。

投资人

谢谢。然后第三个问题，就是我们看到 Anthropic 他们用自己的模型做自己的产品，推出了很多纵向的金融、法律.....甚至未来要往医疗方向走，您觉得在这种纵向应用方面，某个阶段我们也会考虑吗？

梁文锋

我还没有想得很明白，未来我们国内商业模式是怎么样，或者说最顺利的路径是怎样的，我们还没有到那个阶段。国内情况和国外情况不一定一样，国内到时候是什么样，我觉得现在还比较难判断。

国内我们现在以目前的情况来看的话，我觉得最合理的做法应该是全力做通用的 Agent，其他的 Agent 优先级应该更低，包括金融、医生这些 Agent。要先做 Coding，因为 Coding Agent 能够做到很多，还有很多垂直的 Agent。

现阶段我们觉得最重要的，应该还是 Coding Agent。

投资人

讲得很清楚，谢谢您。还有一个问题，我们也想请教您。其实我们做 DeepSeek，也很佩服您，一直在以一个非常纯粹的科研方式来做 DeepSeek。但现在这个行业确实也走到了资本市场，走到了资本化这条路上。您又是一个非常负责任的人，无论是对小伙伴还是对投资人，您也都是非常负责的。

那么，在纯粹科研、纯粹做 AGI 的方向和资本市场之间的平衡，您未来肯定还要走资本市场，肯定还有公众股东，您觉得以后怎么 balance？这个您是怎么考虑的？

梁文锋

我现在觉得应该是能够做到的，就都要。假如说我今年能够有几个亿美金的 B 端收入，再加上我们 C 端有用户，那么这本身就已经有一定的商业基础。以明年我们 B 端有收入，如果这个需求可以再增大的话，公司离净利润已经不远了，可能就已经是净利润了。

可能就已经不是一个纯烧钱的阶段了，所以我感觉后面能做的、能操作的动作应该还是比较大的。或者说，最坏情况卖 API，可能都能够支撑一个上市公司。如果说技术后面没有新的进步了，我们的技术就冻结在这里了，那么最后我们就全力卖 API，把这些服务做好，我觉得也够的。所以我觉得还是有信心的，确实没有那么难。

因为确实是在一个杠杆高的地方，又在一个发展非常快的领域，它可能就是没有那么难。只能讲，我们希望有更大的梦想，但是我们也有保底可以拿出来的业绩。

投资人

谢谢您，讲得很好。问题问完了，谢谢。

感谢您分享。前面提到一些问题，想跟您再问一下。因为我觉得 DeepSeek 其实最大的和其他公司的区别在于，我们的组织跟其他人不一样，或者说组织形式不一样。但组织形式既跟我们的目标相关，可能也要去考虑组织自身的边界和效率。我不知道，从宏观的角度考虑，我们这个组织形式会有一个好的学习对象吗？

历史上可能 Bell Labs，还是一个什么样的形态可能是比较理想的？还是说我们自己觉得也没有理想的，更多还要靠我们逐渐去探索？可能在一个新的时期，只能靠我们自己来做探索。因为我们的组织形式其实跟美国的几家公司肯定都不一样，对吧？三家公司自己也不一样，但他们至少会从一个商业公司的角度出发去探索。

我可能主要想再问一下组织这个问题。

梁文锋

首先，我们没有模仿的对象。每一步都是我们从实际情况出发，实事求是，根据实际情况来做决策，找到我们应该怎么做。所以它是一个时代的产物，或者说是现实情况的一个反应，它并不是一个模仿的结果。就是说，在这个情况下，我确实最优解可能就是这样的，或者说我自己认为，我们自己选的路就是这样。

每一步我们肯定都是思考过的，肯定都是选过，反正选的结果就是这么选的。它并不是因为看到谁这么选，我们这么选，而是因为我们分析了利弊而这么选。那么在未来可能也是一样，我们并没有去模仿谁。

我觉得我们跟 Bell Labs 还是不一样的，因为它明确是不需要有商业化的，因为它是个.....但是我们明确是要有商业化的。我们最终还是要能活下去，我们毕竟是一个公司，政府不会给我一分钱。所以我们可以有非常远大的使命，但我们归根结底是一个公司，我们要考虑怎么活着。

所以 B 端对我们肯定是重要的，因为可能以后可能得靠它活着。只是说现在不重要，因为它现在是个成本线。所以我觉得这个还是不一样的，跟 Bell Labs 还是不一样的。我们本质上还是一个公司。历史上也有很多公司，它也有利润以外的追求，但并不能说它有利润以外的追求，它就不是个公司。

很多公司做得很伟大，因为它有一种利润以外的追求。那个追求最后不但没有影响到它的商业化，反而能让它商业化得更好。我们本质上还是一个公司，只是说我们在考虑赚哪些钱、什么时候赚钱、赚多少钱、靠什么赚钱，只是说我们有取舍。

投资人

梁总，我有两个小问题，快速跟您请教一下。一个是说，您刚刚其实有提到 MILOS 的这个，可能不是一个很确定的目标，但是肯定会往这个方向去做。然后您也提到激活参数可能比如说...下一代可能会是在 150 到 250B 左右的一个情况。这种情况下，您觉得 150 到 250B 是对标 O4.7，还是可能会对标的？这是第一个问题。第二个问题，您前面也提到了，我们现在在整个 inference 上面用了一些除了 CUDA 以外的编译语言。

我理解原来咱们基于英伟达的生态，可能会用 PTX 这些比较多。现在用上更多类似您刚提到 TileLang 这些话，是不是会大幅降低我们在 inference 上面的一些效率，或者短期降低效率？我不知道您会怎么看编译语言的这种变化带来的效率损失，还是说长期其实是一个可以补充的状态，是提高效率？

梁文锋

是提高效率，是大幅提高效率。

投资人

OK，反倒没有什么负面的影响？

梁文锋

梁文锋对，是大幅提高效率。所以说这是一个机遇。相当于以前你是离不开 CUDA 的生态的，现在我们可以抛弃它的生态，用一个更简单的方法，就用 TileLang。它是高级语言，写那个程序也是很快的，要写的代码量很小了，我可以把它重写一遍。

投资人

投资人明白。所以这两个都其实是一个，像您提到的 AI 带来的比较大的机会，不是一个可能短期需要去弥补的缺点。

梁文锋

梁文锋对，它是技术发展的大机会。它不是 AI 的机会，因为我们还有一个项目，我们在用 AI 来写 TileLang。

投资人

那是不是更快？

梁文锋

梁文锋现在所有 TileLang 都是人写的，但是它已经比原来写 CUDA 的要快很多了。

投资人

明白。所以哪怕是对于硬件底层的执行效率，您觉得也是没有影响？

梁文锋

损失 1%到 2%，我觉得是可以接受的。

投资人

OK，明白，理解了。好的，谢谢您，很清晰。

主持人

其他人还有没有问题？没问题，我们今天就先这样。

投资人

好的，谢谢。非常感谢大家的时间，谢谢梁总，谢谢。

主持人

拜拜。

梁文锋

拜拜。