梁文锋投资人会议答问实录

中英对照 · Liang Wenfeng's Answers at the Investor Meeting (Chinese-English Bilingual Edition)

说明：本文整理自网络流传的梁文锋在投资人会议上的答问纪要，共 11 个主题、118 条，已剔除原文末尾的推广内容。英文为对照翻译，仅供研究参考，如有出入以中文原文为准。

01 愿景与克制 / Vision and Restraint

中文原文	English Translation
1. 我们一开始来做这个公司，初衷没有想到说我最后要赚多少钱，要到资本市场上去，要上市，要怎么样的。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。	1. When we first started this company, we never thought about how much money we would eventually make, about going to the capital markets, about an IPO, or anything of the sort. The first few dozen people never thought this way at all. If someone had, they would not have joined.
2. 我们是怀着一个对这个世界非常大的善意来做这个事情，我们觉得这是对人类有用的，这是一个金钱以外的事情。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。	2. We set out to do this with great goodwill toward the world. We believe it is useful to humanity, something beyond money. Our original aspiration and our vision, the vision we hold to this day, were never about maximizing commercial profit.
3. 管理一个大公司，靠的不是你的规章制度，靠的是愿景。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。	3. To run a big company, you rely not on rules and regulations but on vision. A vision is not a slogan hung on the wall. It is about how you act, not how you talk. It is how the company actually runs.
4. 我们是没有组织的，就是愿景驱动的，靠一个愿景来组织。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。	4. We have no formal organization. We are vision driven, organized around a vision. We do not operate by asking what KPI to hit, and there are no performance appraisals. There is only the vision.
5. 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。	5. This vision is not even written down. Nothing has ever been put on paper. It lives in the way we do things and in our attitude toward the world.

中文原文	English Translation
6. 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。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。	6. We do not have many other advantages, no special prowess. We are not richer than others, and our people are not necessarily better than those at other companies, they really are not. When we founded the company two years ago, we had neither much money nor many GPUs, and we had no fame and no pulling power. We are just a group of very ordinary people.
7. 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。	7. The more restraint you show, the easier it may be to succeed. At least so far, this has been borne out, and so far it explains things. Otherwise there is no way to explain why we made it. We had no weapons, our starting point was very low, our resources were very few, and our people are really just a random group of ordinary people.
8. 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面的哪一部分利益、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。	8. AI is too big, and the prize at stake is too big. We stay very restrained. As long as we can pull it off, the eventual rewards will be enormous. Take any small slice of it and the reward is still enormous. So there is no need right now to think about which slice to take or how to take it. I do not think we need to consider that at all, because the prize is big enough.
9. 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。	9. During last year's Spring Festival, our user numbers suddenly surged, but we did not chase retaining those users, or monetizing them, or grabbing commercial gains and cashing in on them. We did not fight for users and we did not go after the money, but we worked very hard to serve our users well.
10. 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。	10. We have never thought about building the next super app, or competing with anyone, or becoming the next ByteDance or the next Tencent. Nothing of the sort. I believe the AGI opportunities ahead are enormous, and they will always be enormous.
11. 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。不开源这个事，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。	11. Restraint is a strategy. Sometimes you can give up one thing in exchange for more of something else. Open sourcing follows the same logic. It can be seen as pressure we put on ourselves, or as profit we give up.

中文原文	English Translation
12. 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。	12. My understanding is that this kind of restraint increases, over the long run, our probability of achieving AGI. When I look at anything, I have no doubt that AGI will have enormous commercial value. On that basis, my first consideration is not how to grab a slightly larger share. My first consideration is how to increase the probability that we succeed.
13. 我们一直非常克制，不愿意跟任何一家互联网大厂或者小厂成为对手。我希望我能够给他赋能，或者希望我能够协助大家去做这个事情，希望能够帮助大家做这个事情。	13. We have always been very restrained, unwilling to become an adversary of any internet company, big or small. I hope we can empower them, assist everyone in doing this, and help everyone get it done.
14. 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。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。	14. We have held this attitude all along, and we have not ended up with anything less because of it. Open sourcing, our goodwill, the help we give others, none of this has cost us anything. If anything, it has added to our gains. It looks counterintuitive, but that is how it is.
15. 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。	15. Our goal is AGI, but we have been commercializing all along, which is why we have consumer users and enterprise revenue. Judging by experience so far, this strategy has worked.

02 AGI 路线图 / The AGI Roadmap

中文原文	English Translation
16. 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，给它完整的指令。	16. If you can describe a problem clearly and give it complete context and instructions, it already surpasses humans. But there is a definition here, a premise. You must give it complete context and complete instructions.
17. AI 并不能够替代你的员工。但如果说 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，那么就可以替代天下的人了，所以我们离下一步还差一个持续学习。	17. AI cannot yet replace your employees. But if AI had the ability to learn continuously, the way a new hire spends two months learning at your company, then it could replace anyone under the sun. So the missing step before the next stage is continual learning.

中文原文	English Translation
18. AI 的发展，我们可以理解成它是一个阶梯。去年走的阶梯是思维链。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。	18. We can think of AI's development as a ladder. Last year's rung was chain of thought, because we found that chain of thought could take intelligence to a higher level.
19. 今年的阶梯就是 Agent，因为我们发现，用 Agent 的方式，即便多些事情也可以做，它的能力范围会更大，它的智能上限会更高。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。	19. This year's rung is agents, because we found that with agents you can do considerably more things. The range of capability is wider, and the ceiling of intelligence is higher. Agents use CoT, and CoT in turn builds on the previous rung, the language model. So no step along the way was wasted.
20. 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。	20. After agents, the problem we think must be solved is continual learning, how to let a model keep learning. Not through one more heavy training run, but the way a person learns continuously over a long period of time.
21. 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。	21. After continual learning, we may arrive at a singularity. At that point, once the model can learn continuously, it can already do everything humans can do. It can develop its own next version, conduct its own research, and build ever more advanced AI models.
22. 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。	22. This singularity is not really a singularity. It too is a gradual process, probably a fairly long one, not an abrupt jump. But out of habit, we all tend to think of it as a singularity.
23. 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。	23. This is our conjecture, the timeline as we see it. First solve continual learning, then reach the singularity of intelligence, the self-iterating singularity, and only then embodied intelligence. Once it is embodied, it walks into the real world. It can do your housework and take care of you in old age.
24. 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。	24. If we first solve continual learning, then the self-iterating singularity, and then embodied intelligence, the road becomes much easier, because further along you can use the earlier technology to help develop the later technology.

中文原文	English Translation
25. 我们只做 AGI 的主线。AI 领域很广泛，有很多东西我们觉得它不在这个主线上面，比如说 3D、视频生成，我觉得可能跟智能的主线没有太大的关系，我们不会去做。	25. We only work on the main line toward AGI. AI is a broad field, and many things are not on that main line, 3D and video generation for example. I think they have little to do with the main line of intelligence, so we will not do them.
26. 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。	26. When video generation first appeared it was all the rage, as if it were compulsory, as if you were not an AI company unless you did it. I find that strange. If you think it through carefully, it has nothing to do with the roadmap of intelligence.
27. 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。我们不会因为它是一个好商业而去做它，我们只会因为它是智能路线图上的东西，才会去做。	27. Commercially, it is a good business, no doubt. But it has nothing to do with intelligence. We will not do something because it is good business. We will only do it because it sits on the roadmap of intelligence.
28. 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。	28. In our judgment, world models and intelligence are not the most important things at this stage. What matters most is AI training itself, and how to solve continual learning after training. That is our company's judgment. Of course, every company judges differently.
29. 我们现在比较相信一个叙事是，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。	29. One narrative we now believe in is that AI can accelerate AI research. In other words, progress is not linear, because you can use AI to speed up your own research. So further down the road it may be nonlinear.
30. 我觉得具身肯定还是要进入，最终具身。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。	30. I do think embodiment is where things ultimately end up. In the end it must be embodied. For an ordinary person, what he needs is not a computer, right? Ordinary people eat, drink, play, dress, live and travel. They do not need a computer. What they need is embodied intelligence to meet concrete, physical needs.
31. 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。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。	31. What do we hope AGI can do? Help us iterate the next model version. And once we have embodiment, we hope for the same thing. Let it iterate the next generation of the embodied system. Let it build the next robot.

中文原文	English Translation
32. 下一代模型的核心能力，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。	32. The core capability of the next generation model must be continual learning. Only then does it deserve the name. Before that, all we can do is cut costs, improve quality and increase speed. But a real breakthrough will require continual learning.
33. 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。	33. Today's agents are limited because they cannot learn continuously, at least not effectively. If continual learning is solved first, AI's capability will be extremely strong, and it will greatly raise the efficiency of our own research.
34. 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。	34. If continual learning comes first, general intelligence may become easy, because you use it to do the work. That is the outcome we hope for. It saves us effort and makes life easier. Otherwise, building general intelligence by hand is tiring, bitter work, data intensive and labor intensive, with poor returns for the effort.

03 团队与人才 / Team and Talent

中文原文	English Translation
35. 我们前面的经历给我的启示是，AGI 这个愿景是很强大的。这个人才优势不是说我的人比他更聪明，而是这些人才我怎么组织起来，怎么激励他，然后怎么合作。	35. What our past experience taught me is that the vision of AGI is very powerful. Our talent advantage is not that my people are smarter than his. It is about how I organize these people, how I motivate them, and how they work together.
36. 你把聪明的人聚在一起，并不是说他自然而然就能够合作，自然而然就能够非常有激情地去奔一个目标、去完成的，所以你需要一个愿景。	36. Put smart people together and they will not automatically cooperate, nor automatically charge with passion toward a shared goal. That is why you need a vision.
37. 我们最大的核心利益是要保持团队的稳定性。这是我们最大的核心利益，甚至可以认为是唯一的核心利益。只要我能够保持团队的稳定性，我一定能做成，一定能做成 AGI，就这么简单。	37. Our greatest core interest is keeping the team stable. It is our greatest core interest, arguably our only one. As long as I can keep the team stable, we will definitely make it, we will definitely achieve AGI. It is that simple.

中文原文	English Translation
38. 钱肯定不是问题，资源不是问题，其他要素都是容易获得的。对我们来讲，只有一个核心利益，只有一个没法退让的：我们必须要保持团队的稳定性。	38. Money is certainly not a problem, resources are not a problem, and every other factor is easy to obtain. For us there is only one core interest, only one thing on which we cannot yield. We must keep the team stable.
39. 这也是我们面临的一个非常大的挑战，或者说，我觉得是最大的风险。当然，这个风险随着我们近期的这一次融资，得到了比较大的解除。因为大家拿到的期权都还是比较多的，金额还是比较大的。	39. This is also a very big challenge for us. I would say it is the biggest risk. Of course, our recent round of financing has largely defused that risk, because the options everyone received are fairly substantial, a fairly large amount.
40. 从团队稳定性来讲，只要最重要的一些员工、最老的员工能够稳定，那么其他人是不太会走的。其他人哪怕期权少一点、收入少一点，他也不会走。因为他不是全奔着钱来的，大家都希望在一个能够做成 AGI 的环境里面去做这个事情。	40. In terms of stability, as long as the most important and the longest serving employees stay put, the rest are unlikely to leave. Even with fewer options or lower pay, they will not leave, because money is not all they came for. Everyone wants to do this work in an environment where AGI can actually be achieved.
41. 其他都是时间问题，其他最多导致我们晚半年、晚一年，但是不会说做不出来。肯定是不缺钱，肯定是不缺资源，其实这些都是不缺的。	41. Everything else is a matter of time. At worst it delays us by half a year or a year, but it will not stop us from getting there. We are definitely not short of money, definitely not short of resources. None of these are lacking.
42. 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。	42. Our gap with the US is mainly in resources. The gap in people is not large, almost none at all, because it is the same pool of people, Chinese people. Among the Chinese, some stay in China and some go abroad, but it is not the case that the smart ones all go abroad. Not at all.
43. 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。	43. Talent is not the bottleneck. Resources are the biggest bottleneck. Resources affect talent development first. With little compute, we get fewer chances to run experiments, so overall our talent lags behind America's. The talent gap, at its root, is also a compute gap.
44. 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。	44. The AI talent shortage is also a phase, and we have already seen it ease significantly. There is really no shortage of AI people. Every company trains people up quickly. Training people is fast.

中文原文	English Translation
45. 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。	45. There are somewhat too many companies in China building models now, still too many. The US has maybe three. Too many players in China are building foundation models. In the end there will not be a need for so many. It is bound to consolidate.
46. 我们公司的管理其实是两条线：一条线是从上到下，一条是从下到上。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。	46. Our management actually runs along two lines, one top down and one bottom up. Bottom up means everyone works on what they want to do. Nobody manages them, and there are no KPIs.
47. 一般我们希望员工还有一半的时间，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。	47. Generally we want employees to have half their time unscheduled, free to do whatever they want. It is a space for research, where they can explore on their own, pursuing whatever they judge to be important, with no preset requirements.
48. 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。	48. We rarely work overtime. There are two reasons. First, research needs a relaxed environment. If you squeeze people too hard, they cannot do research. Since the work requires your own interest, requires you to think about these problems on your own, only a relaxed environment makes exploration possible.
49. 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。	49. Second, we are extremely focused. Extreme focus means there are very few things to do. With fewer things to do, there is no need for overtime. This comes from the same root as the restraint I talked about.
50. 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。	50. Our company as a whole is built on consensus. I do not decide everything by myself. I seek consensus. My authority and influence inside the company are built on consensus.
51. 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。	51. Our decision mechanism is really a consensus seeking mechanism. It is not that I can push something through on my own. Only when there is consensus can it move forward, and only then do I push.
52. 随着人员的增加，我们会做这个调整。应该是我们马上就得做这个调整，因为我已经在做这个调整。已经不做这个调整的话，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。	52. As headcount grows, we will make adjustments. In fact we must make them right away, and I have already started. Without them, many things cannot move forward. There really are many departments that should have an organizational structure.

04 算力与资源 / Compute and Resources

中文原文	English Translation
53. 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。	53. How many GPUs do we need? Right now, the more the better. Within what we can afford, more GPUs are always better, no doubt about it. So our current strategy is, at a reasonable price, buy as many GPUs as we can.
54. 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。	54. In reality, spending all this money is very hard. You cannot buy that many GPUs. They are hard to get and expensive, and we cannot pay any absurd price. We still have to make sure the price is reasonable. If we manage to spend 20 billion yuan this year, our procurement team will have done a superlative job.
55. 我们跟美国之间最大的差距是在资源上面。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。	55. The biggest gap between us and the US is in resources. On compute, for one thing GPUs are simply unavailable in China. For another, our capital investment is smaller than America's, quite a lot smaller. Talent compensation accounts for a very low share of that investment. You see them offering salaries of a hundred million dollars, but do the math. Talent pay is still a small share. The bulk is compute.
56. 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。	56. All the differences we see, in talent, in model capability, in applications, can be traced back to differences in compute resources.
57. 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。	57. Our gap with the US might be twelve months behind, maybe twelve to eighteen, or six to twelve. Put simply, about two years behind, while getting the job done with one twentieth of America's compute.
58. 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。	58. Today's narrative is, one to two years behind, but on one twentieth of their compute. Going forward we want to rewrite that narrative. Still a fraction of their compute, but with the time gap compressed much further, to six months, to three months. I think that is a goal.

中文原文	English Translation
59. 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。	59. We believe in scaling. Bigger scale definitely means better results and unlocks more capabilities. What stops us from scaling is compute. It is not that we do not want to scale. We simply do not have enough compute to do it.
60. 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、能够训练的模型是多大，是这样算出来的，并不是这个模型就够了。	60. We train a model of this size not because I think this size is enough, but because this happens to be what our resources allow. I work backward from my resources to the largest model I can afford to train. That is how the size is determined. It does not mean the model is big enough.
61. 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。	61. When Silicon Valley says scaling has hit its limit, that is true for Silicon Valley. For the Chinese, we are still far from that point. We have never scaled to that degree at all. And scaling includes the scaling of data, of model size, and of training cost.

05 国产芯片与生态 / Domestic Chips and the Ecosystem

中文原文	English Translation
62. 英伟达 CUDA 的护城河在快速地被瓦解。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。	62. NVIDIA's CUDA moat is being breached quickly. For one thing, we now have AI, and with AI, building such an ecosystem is much easier than before, because AI can write code.
63. 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。	63. The market for compute cards is now bigger than the market for gaming cards, so there is no reason the two must remain coupled. The trend now is decoupling. Going forward it is all dedicated silicon, whether from Huawei or from NVIDIA itself. None of it will look like what came before.
64. 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，我们能够看到有一个事情被验证：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。	64. Domestic AI chip substitution now faces a historic opportunity. We believe that within a year, one thing will be proven, that the domestic chip ecosystem is entirely viable. People used to think it was not, that domestic chips could not be put to use or were hard to use. Within the next year, I believe, we will reverse that perception, or the facts will reverse it.

中文原文	English Translation
65. 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。	65. Domestic AI chips have no problem in hardware or in ecosystem. The only problem is insufficient capacity. Adapting to domestic cards presents no obstacle, and NVIDIA cannot stop it. In a normal commercial environment where I could buy NVIDIA GPUs, domestic substitution would be relatively hard. But when NVIDIA GPUs cannot be bought, everyone is forced to go domestic.
66. 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。	66. When V3 was trained, it still used NVIDIA GPUs, but it no longer used NVIDIA's ecosystem. V3 ran on NVIDIA hardware without NVIDIA's ecosystem. We first wrote a high level compiler called TileLang, and then did everything else on top of the TileLang ecosystem, so we barely depend on NVIDIA's ecosystem at all.
67. 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。	67. I am quite optimistic about domestic compute. On this front, I think NVIDIA is digging its own grave. Huawei's SuperPoD, the Huawei 950 SuperPoD, can fully substitute for NVIDIA's GB200 and GB300 on both performance and price.
68. 四张华为卡顶一张英伟达的卡。	68. Four Huawei cards match one NVIDIA card.
69. 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。	69. As for our chip gap with the US, I believe there will be no ecosystem gap going forward, but on the chips themselves the gap is fourfold plus two years.
70. 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。	70. Right now we mainly work with Huawei. Huawei does the adaptation themselves, but we take part in the ecosystem and get deeply involved in it. Huawei's problem is still insufficient capacity.
71. 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。	71. I do not quite believe that five years from now we will still be stuck on capacity. We are definitely stuck on it now, and this year, next year and the year after we may still be stuck on it. But five years out, maybe not. I am fairly optimistic.

06 竞争格局与行业判断 / Competitive Landscape and Industry Outlook

中文原文	English Translation
72. 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。	72. The final performance gaps between models should be understood in combined terms. Comparing model performance is only meaningful at equal cost, just as you compare two cars at the same price point.
73. Anthropic 跟现在超过 OpenAI，这是不是长期的？我觉得这不是长期的，这肯定是阶段性的。OpenAI 和 Google，未来大概率还是会交替上升。	73. Anthropic has now overtaken OpenAI. Is that long term? I do not think so. It is certainly a phase. OpenAI and Google will most likely keep trading the lead going forward.
74. 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。	74. When the global division of labor in AI settles, the role Chinese companies will most likely play is still that of the largest producer. By common sense, we have the largest capacity, including chips, where our capacity may also end up the largest, and we have the most electricity.
75. 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产并没有太大的区别。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。	75. The Chinese will make this product the cheapest. And in terms of quality, after all, for many goods today, made in China and made in America are not very different. AI will probably be the same in the future, except that Chinese made AI may be cheaper. The cheapness may be systematically low, the same way Chinese services in other industries are cheaper.
76. 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。	76. The final gaps should come down to three things, cost, time, and user experience. Beyond those, there may be no gap at all.
77. 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。	77. Cost is certainly one difference. I would say cost ranks first. Second is time, when you can deliver. A few months earlier or later makes a real difference.
78. 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。	78. OpenAI believed from day one that it could truly monopolize this world, but in reality it will meet many, many challengers. With challenges come hard times. America will be challenged, and in the future it may face a challenge from China, because the Chinese are willing to take less and can still provide the service.

中文原文	English Translation
79. 拿得多的人会被拿得少的人打败。甚至你还不用真的拿得多，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，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。	79. Those who take more will be beaten by those who take less. You do not even have to actually take more. If your vision is to take more, you will be beaten by someone whose vision is to take less. In reality nobody has taken the money yet. It is only a vision. But if your vision is to take more, you have already lost, and you will face greater difficulties.
80. 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。	80. For us, the point is not to take the maximum profit, or to price for revenue maximization, but to earn only a reasonable return. That is the explanation. I genuinely believe this. I am not looking for excuses, because there is no need for excuses.
81. 我觉得在很多体验方面，有可能我们是能比美国做得好的。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，所以中国还是会有竞争力的。	81. In many aspects of the experience, I think we may do better than the US. On the product side, our product capability is not necessarily worse than America's. Costs should also be lower. So China will remain competitive.
82. 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。	82. The cost part is easy to understand. They do not have to do it, so they never develop the capability. They certainly do not take it as seriously as we do. We can treat it as something extremely important. For them, it is not important.
83. 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。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。	83. For large models, maybe two big companies and two small ones are already enough. The gaps come down to only two things, time and cost. So no player will make windfall profits. I do not think there will be windfall profits. Whoever controls costs well earns a bit more, whoever controls them poorly earns a bit less. That is all.

07 模型研发与技术 / Model R&D and Technology

中文原文	English Translation
84. 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。	84. Half of our company, perhaps, would say in private that OpenAI is better. Anthropic has a first mover advantage, but that advantage should fade quickly. It is not something it can hold for the long term. All three players are formidable. Among the three, our efficiency is the highest. What we spend, what we burn, should be the least.

中文原文	English Translation
85. 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。	85. We have been working on multimodality all along. For the product it matters a lot. For consumer facing products it matters a lot. But for the ceiling of intelligence, it is a component, not the main line itself.
86. 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。	86. We should release the corresponding models. V4 and its follow on versions will support native multimodality. But for us, for intelligence, multimodality is a component. We do not treat it as intelligence itself.
87. 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。	87. All I can say is that for language model scaling, I see no ceiling for now. Neither at our current level of intelligence nor at the level in the US do I see a ceiling.
88. 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。	88. Many of us internally think this way. It must first be useful to ourselves, first serve our own use, because this is the fastest way to AGI. If it works well for us, it probably works well for others too, but first it must work well for us.
89. 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。	89. The first goal of the models we build is not that everyone else finds them useful, but that we find them useful. They must first be useful to us. Once they are useful to us, we can develop the next version faster.
90. 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。	90. We call this a lucky draw. The barrier to entry is low. Anyone can take a draw. But what you draw out of it, whether that depends on talent or on something else, I cannot say. So this is not something we need to allocate resources to. Where we differ from other companies is that we spend time discussing and thinking about this question, and we treat it as important.

08 商业化与定价 / Commercialization and Pricing

中文原文	English Translation
91. 我们的 API 定价是一个合理的利润，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。	91. Our API pricing targets a reasonable profit, roughly the level where buying a batch of equipment on the market pays for itself in ten months. I consider that a reasonable profit.

中文原文	English Translation
92. 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。	92. If we wanted to maximize profit, we should price higher, because in this price range user demand is inelastic. Raise the price by half, or even double it, and token consumption barely changes.
93. 我们的一个模型，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。	93. With one of our models, we worried demand would be too high, so we set the price relatively high at first, and the team was not happy about it. Later I cut the price to a quarter of that, and everyone was delighted.
94. 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。	94. The ceiling of the enterprise business should still be demand. Given this generation of AGI and AI technology, enterprise demand should be finite. It will grow fast, but it is not infinite. In the end it is constrained by demand, not by compute.
95. 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。	95. I now think we can do both. Suppose this year we can book a few hundred million dollars of enterprise revenue, and on top of that we have consumer users. That in itself is already a commercial base. If next year our enterprise revenue grows as demand expands, the company will not be far from net profit. We may already be profitable.
96. 最坏情况卖 API，可能都能够支撑一个上市公司。就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。	96. In the worst case, selling APIs could probably still support a listed company. If technology stopped advancing from here and our technology froze at this point, we would go all in on selling APIs and do those services well. I think that would be enough.
97. 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。	97. As things stand, I think the most sensible approach is to go all in on general purpose agents. Other agents take lower priority, including finance and medical agents. Coding comes first, because a coding agent can already do a great deal, and there are many vertical agents beyond it. At this stage, the most important should still be the coding agent.
98. 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。	98. I see low cost first of all as a result. Our models have consistently moved toward lower cost at the architecture level, and that is tied to our vision. We also have many algorithmic methods that can push costs lower still.

中文原文	English Translation
99. 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。	99. There is another reason to push costs down. The lower the cost, the bigger the model I can train, the bigger the model I can afford. With the same compute, with compute being limited, higher efficiency lets me afford a bigger model.

09 开源策略 / Open-Source Strategy

中文原文	English Translation
100. 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。	100. I think we will open source, and our strongest models will probably be open sourced too, because I cannot see what good closed source does, any necessary good. ByteDance's models are closed source. What good does that do them? I cannot see any.
101. 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，也很难很难，没有那么容易。	101. Even if you open source a model and tell everyone everything, the bar remains very high. For others to actually put it to use, the bar is very high. Using it at all is hard, and using it at very low cost is even harder. It is not that easy.
102. 开源并不会影响收入。开源，我觉得对我们的商业模式是没有任何影响的。	102. Open sourcing does not affect revenue. I believe open sourcing has no impact at all on our business model.
103. 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。我们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。	103. Nor do I worry about others deploying our models and competing with us, not in the slightest. In fact we hope they get them deployed. We help the open source community as much as we can, assisting everyone in deploying our models.
104. 我们在对外面打交道的时候，我们的态度是：我们只做 AGI 的主线。在对外面打交道的时候，我们是很愿意协助、帮助任何人，甚至我们的竞争对手，包括阿里、智谱、月之暗面，去做得更好。因为我们并不损失什么东西，我们本来也是开源的。	104. In dealing with the outside world, our stance is that we only work on the main line to AGI. And we are very willing to assist and help anyone, even our competitors, including Alibaba, Zhipu and Moonshot AI, to do better, because we lose nothing by it. We are open source anyway.

中文原文	English Translation
105. 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。	105. Are the models we open source the same as the ones we deploy ourselves? They are the same. We would never open source a weaker model while deploying a better one ourselves. No, they are the same.

10 数据与后训练 / Data and Post-Training

中文原文	English Translation
106. 数据应该几乎就等于模型的一半。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。	106. Data is almost half of a model. And before that comes the question of labeling. On data annotation, it comes down to our capital structure. With the way our capital is invested, we cannot afford the cost of that much high quality annotation, because it is expensive.
107. 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有成本优势，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。	107. The cost of data annotation in the US is not much different from China. Annotating data in China brings no cost advantage, especially for high end data, which makes it hard for us to invest in annotation the way the US does. This path is very hard in China, because annotation is simply too expensive. Whether we outsource it or do it in house, it hurts either way.
108. 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。	108. Right now we basically walk on two legs. It is not that we cannot annotate at all. Some annotation is cheap and some is expensive. We do the cheap kind first.
109. 你也可以认为，现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。	109. You could say that half of our company is now annotating data. Half of our core researchers, our most important people, are annotating data. We are concentrating on it. At this stage, solving the AI problem comes down to annotating data.
110. 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。	110. The bottleneck in high quality annotation, I think, is time. It simply takes time. OpenAI, the overseas players, Anthropic, they all started earlier, with more capital and more GPUs.

中文原文	English Translation
111. 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。	111. Hallucination in large models hurts the user experience quite a bit. There is a way to solve it, but it is a long term project. Hallucination can be seen as something solvable through better post-training, a problem that can be fixed, or at least improved.

11 组织与公司定位 / Organization and Company Positioning

中文原文	English Translation
112. 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。	112. First of all, we have no model to imitate. Every step starts from actual conditions. We seek truth from facts, make decisions based on reality, and work out what we ought to do. So it is a product of its time, or rather a response to actual circumstances, not the result of imitation.
113. 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。	113. We are clear that commercialization is a must. In the end we have to survive. We are, after all, a company, and the government will not give us a penny.
114. 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，我们有取舍。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。	114. We are at bottom still a company. It is just that we pick and choose which money to earn, when to earn it, how much to earn, and on what to earn it. Many companies became great because they pursued something beyond profit, and that pursuit, far from hurting their commercialization, made them commercialize better.
115. 对于合作伙伴，其实我们这个融资是精心挑选的。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，因为我们还是损害了很多其他人的利益的。	115. As for our partners, this round of financing was in fact carefully chosen. First and foremost, I felt our interests were well aligned, those whose interests align best with ours, who bear us the least hostility, or who most want to see us succeed. Not everyone wants us to succeed, because we have harmed the interests of quite a few others.
116. AI 现在不缺品位和直觉，它缺的是持续学习的能力。AI 的品位和直觉没有问题。你让它写个文章，它的品位和直觉，我觉得没有什么问题。	116. What AI lacks today is not taste or intuition. What it lacks is the ability to learn continuously. AI's taste and intuition are fine. Ask it to write an article and its taste and intuition are, in my view, perfecty fine.

中文原文	English Translation
117. 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。	117. We want to do just one piece of it. AI is too big a thing, it does not need me to... I only do one piece. With focus, and since I believe the prize here is already big enough. If the AI era will produce many trillion scale companies, I believe we will be one of them.
118. 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。但至少这里边是没有利益冲突的，我们是希望合作共赢的。	118. We hope to support more people, but we do not have that much bandwidth. The willingness is there, and there would be no conflict of interest, but whether we actually do it is another matter. At the very least, there is no conflict of interest here. We hope for win-win cooperation.