Liang Wenfeng Investor Exchange Meeting – Recording Translated Into Text

Compiled by: Pang Tong (WeChat: pongtom)

This article was compiled by **Pang Tong using AI**

Note: This draft was automatically transcribed from speech and edited by AI without speaker identification. Brackets indicate audio time positions; some proper nouns and numbers may have recognition errors—refer to the original recording for accuracy.

Content Timeline

【00:00:01】

The same applies to our other colleagues at the company: when we first established this enterprise, our initial motivation was entirely different—we never considered how much profit we would ultimately achieve, nor did we plan to go public or list on a stock exchange. None of us had such intentions from the outset; none of the original dozen or so team members ever thought about it. Had they done so, they wouldn't have joined.

Overall, we approached this endeavor with profound goodwill toward the world, believing it to be beneficial for humanity—an endeavor beyond mere financial gain. Of course, once the benefits became evident, additional temptations emerged; that's a separate matter altogether.However, our original intention, our vision, and the vision we have maintained to this day were not pursued with the sole aim of maximizing commercial profits. I believe this point is particularly crucial.

About twenty years ago, my greatest admiration in management was for Jack Welch, the former CEO of GE. Looking back now, much of what he said may no longer hold true, but one key point he made was absolutely correct: a company's most crucial element is its vision. Managing a large corporation relies not on rules and regulations, but on vision. What is a vision? It's not just a slogan on the wall—it's about how you actually act, not how you talk about it.

I've forgotten Jack Welch's exact words, but it roughly meant this: How do we manage such a large team and organize it? In fact, we have no formal organization—we are driven solely by a vision. This approach has both advantages and disadvantages; in the future, we'll strive to leverage our strengths while addressing weaknesses, but this remains our distinctive characteristic. We don't operate based on specific KPIs or performance metrics—our entire approach is guided by a shared vision.

This vision is not even written down; nothing has ever been formally documented. It resides in the way we approach our work and our attitude toward the world. While each individual in our company may have a unique understanding of this vision, and personal interpretations may vary, there is overall alignment on a fundamental level. I believe it stems from a profound goodwill toward the world and a desire to make a positive impact. This principle serves as the foundation for how we operate as a team.

First, I'll outline the main points; after that, you may ask questions. My subsequent discussion will revolve around this vision. This vision is genuine—it's not fabricated; we truly believe in it and act accordingly. Without it, we couldn't explain many of our practices. Why do we insist on open source? Because the vision itself demands open source. Without this vision, it's impossible to unite people effectively. For example, ZhiPu also adopts an open-source approach, but its model differs from ours.

Zhipu's decision to adopt open source felt somewhat forced—they didn't believe it was their original intention—but for us, it was precisely that. From the outset, we had a clear vision regarding open source: first, a clear strategic direction; second, our conviction that open source would facilitate the commercialization of AI. This may sound contradictory or counterintuitive, given that historically open source and commercialization have often been at odds.

But I believe AI is fundamentally different from the past. Historically, a software company's annual market revenue might have been in the billions of dollars; once it exits the open-source ecosystem, that value could plummet to just tens or hundreds of millions. In contrast, AI represents an enormous scale—ultimately potentially accounting for as much as 10% of humanity's GDP, for instance. That's truly a staggering figure.

No one can monopolize this endeavor alone—you must share it with others; otherwise, you simply won't survive. This differs from developing open-source software, where the market potential is relatively limited. However, AI represents a truly massive field. Those who attempt to claim exclusive benefits will inevitably be left behind by history. This is fundamentally an objective law and a matter of historical perspective.

Just because I don't adopt open-source practices doesn't mean I can monopolize this market; that's theoretically inconsistent with reality. You'll inevitably face significant resistance and other obstacles to achieving your goal. In such circumstances, I believe it's not necessary to adhere strictly to traditional business logic. You need a mechanism to ensure that the benefits you gain are limited—only then can you succeed. Self-restraint is essential, in my view.

If we want to successfully implement AI in our hands, first and foremost, I believe restraint is essential. We must not think that a certain percentage of humanity's GDP belongs to us, or even a specific percentage of China's GDP. The more you hold such thoughts, the less likely you are to succeed. Therefore, from the very beginning, we have emphasized the need for restraint. The more restrained you are, the greater your chances of achieving this goal. This is a commercial consideration, though it is undoubtedly a macro-level one.

I think this aligns with my intuition—at least, that's what I genuinely believe. We don't possess any significant competitive advantages; we lack exceptional capabilities, aren't wealthier than our competitors, and our workforce isn't superior to that of other companies either—in fact, none of these claims hold true.You see, when we founded this company two years ago, we had neither substantial funds nor numerous business cards, no notable reputation, nor any influence—we were simply a group of ordinary people.

【00:11:49】

I am truly just an ordinary person. If I were to describe a favorite narrative as one where ordinary people achieve extraordinary things—rather than geniuses doing so—it stems directly from our restraint, aligning perfectly with both our self-discipline and vision. So, could open-source development conflict with commercialization? In the realm of AI, I believe that without restraint, progress is impossible.

Open source is one aspect of restraint; indeed, our restraint manifests not only in open source but across many other areas as well. Yet overall, we need not dwell on the concept of open source or restraint itself. The more restrained you are, the more likely you are to succeed—or at least, that's what has proven true so far and remains logically sound.

Otherwise, there would be no way to explain how we succeeded: we had no advanced tools, started from a very low base with scarce resources, and our team consisted of nothing more than ordinary individuals. As for me personally, I'm just a university graduate—not even from the top-tier institution. This level of restraint is integral to our vision. The field of AI is vast and carries immense potential benefits; therefore, we exercised strict self-restraint, knowing that success would ultimately yield tremendous rewards.

Any modest share you take will yield substantial benefits, so there's no need to worry about which portion to claim or how to obtain it—this consideration is entirely unnecessary given the enormous potential returns. Even a small share is more than sufficient. As we've emphasized before, our focus is on securing a reasonable profit based solely on your willingness, not on the magnitude of profits; this approach differs fundamentally from our API pricing model.

We consider a reasonable profit margin for our API pricing: purchasing equipment from the market and recovering costs within approximately ten months. Given current circumstances—including associated risks and initial investments—we adopt a three-to five-year financial amortization period for servers, yet believe a ten-month cost recovery period is sufficient commercially. Therefore, we deem this approach adequate.

This is our current pricing strategy for the API. For both our V3.2 Flash version and other versions, we recoup the device costs within ten months – this is our standard approach. It's not designed to maximize profits; if that were the goal, prices should be set higher. Within this price range, user demand is inelastic: even if prices are halved or doubled, token consumption remains largely unchanged.

If I double the price, my total revenue would nearly double. Wait a moment—I'll check that. Oh, that's great! Let me tell you a story about our DDCP model. Initially, we were concerned about excessive demand, so we set a high price, which displeased the team. Later, I lowered the price to one-quarter of the original amount, and everyone was delighted. I believe this truly reflects our genuine intentions.

This aligns with the vision I mentioned earlier: we aim for this product to be truly beneficial to users, not merely a source of maximum profit. Instead, we strive to ensure it remains affordable for everyone while generating reasonable profits. I recall that when we implemented price reductions, many colleagues in our company's team celebrated enthusiastically − everyone was delighted. This was precisely the outcome we had worked so hard and meticulously to achieve with this model.

The goal is to offer a product that is extremely affordable yet highly effective, making it accessible to everyone. We find this truly rewarding—it drives us forward, represents our vision, and embodies the shared commitment within our company to achieve this objective. This aspect is particularly unique, as price reductions would undoubtedly not bode well for our competitors; they certainly wouldn't welcome such a move.

If your revenue or ARR is cut in half, your ARR will drop correspondingly. That's precisely what sets us apart—we consider this level sufficient. Internally, we're already highly satisfied with recovering our costs within ten months; externally, we believe this price point is widely welcomed and beneficial for all parties involved—a win-win situation for the company, society, and everyone involved.

I agree that someone just commented on the screen: "A payback period of ten months represents an excessively high profit margin." Indeed, there is still significant room for price reductions and model optimization, indicating substantial potential for overall cost reduction. However, achieving a payback period of ten months is something we can accomplish alone—something competitors cannot. Companies like Alibaba or Tencent, for instance, operate under different optimization frameworks and typically face costs several times higher than ours. There remains considerable room for further optimization.

The reason we haven't continued to lower prices is that there's no room for further reduction. Even if I cut prices further, demand won't increase significantly; or even if I do lower prices, the increase in demand will be minimal. This price point is affordable for everyone, and people are generally satisfied with it—they wouldn't avoid using the product simply because it's expensive. Therefore, reducing prices would not generate additional revenue for the company nor create significant value for society, as everyone is already content with this price level.

No matter how low your prices are, they don't significantly enhance people's sense of happiness in society. That's right—particularly regarding pricing strategies, we definitely shouldn' t adopt a simplistic approach.

【00:25:32】

Maximizing revenue or profit is not our primary objective. This reflects a principle of restraint: while higher prices may boost short-term income, their long-term impact remains uncertain. To me, restraint is a strategic approach—one that involves sacrificing certain aspects to gain greater benefits elsewhere. The decision not to open-source our software follows the same logic; it can be seen either as pressure or as a form of concession.

First, we are delighted with this benefit initiative internally; all employees feel satisfied and achieve a strong sense of accomplishment, which enhances team cohesion. Moreover, it benefits society as well − the public and industry peers alike will find it positive. I believe this measured approach will ultimately increase our chances of achieving AGI's vision. When evaluating this matter, I have no doubt that AGI holds significant commercial value.

Based on this foundation, my priority is not how to secure more market share or capture additional shares; rather, it is how to increase the likelihood of achieving success. This principle of restraint manifests in many other aspects as well. For instance, during last year's Spring Festival, our user base surged dramatically, yet we did not pursue retaining these users, monetizing them, or competing for commercial benefits derived from them.

We don't compete for users or pursue profit-driven goals; instead, we strive diligently to provide excellent user services. We never harbor the ambition of creating the next "super app," competing with others, or becoming the next ByteDance or Tencent—such aspirations are entirely absent from our mindset. While we could have pursued that approach, we chose not to. In my view, this reflects a fundamental principle of restraint.

Don't expect to profit from everything. Once you've acquired users, it seems like you can create the next ByteDance-like company and then dominate the market. I believe this approach is commercially viable and feasible. Last year, we invested heavily competing with ByteDance for users − that was one strategy. However, we chose a more restrained approach: we decided not to compete directly, because what lies ahead may be a "watermelon" (a massive opportunity), while what comes before might merely be "sesame seeds" (small but valuable opportunities). We shouldn't try to capture every single sesame seed.

Admittedly, this may seem like a minor issue at first glance, but I believe the challenges ahead in AI will far outweigh these initial obstacles. Looking back, our decision not to focus on the consumer market last year was likely the right one—because we recognized there was truly a much larger opportunity ahead, while what we tackled initially was merely a small part of the bigger picture. Had I had substantial resources last year and tried to scale this endeavor extensively, what benefits would I have reaped? Nothing at all.

These are my genuine views, as I believe the potential opportunities in AGI will always be immense. We don't even need to consider what position we'll occupy or what our business model will be—those are completely irrelevant. Given such vast commercial prospects, there's bound to be a way forward.

As for the minor details mentioned earlier, we do collect them too, but only in a casual manner—without deliberately pausing to address them or treating them as significant matters. Therefore, metrics like C-end daily active users last year were probably considered minor issues by me; yet we still collected them and maintained user engagement at relatively low cost, anticipating potential future value. Although we currently don't know exactly what these users will contribute, this represents a pure cost expenditure now, but it could prove beneficial in the long run.

Since these products are readily available, we'll naturally seize the opportunity. This year, there's a strong chance that our ARR revenue in the API or AI sectors could also grow significantly. If demand continues to expand and more graphics cards become available, reaching several hundred million US dollars in ARR is entirely feasible.

If AI revenue reaches one billion US dollars, our company's cash flow would essentially become positive, covering all R&D and operational expenses. While this is feasible, we don't consider it a top priority. We'll pursue it, but I believe it's not our primary focus or what truly matters to us at present.

The greater opportunities lie ahead, including last year's consumer-oriented initiatives and this year's business-oriented efforts. I believe these are worthwhile endeavors that need to be executed well—but they aren't our primary focus. In fact, most of our team members don't consider this a particularly critical matter, nor do they view it as equally important compared to AGI development. Regarding open source, we could elaborate further, given that many previous discussions have centered on this topic.

First, I believe we will make our code open source, and likely even our most advanced models. After all, I see no real benefits to maintaining closed-source code—no inherent advantages. Take ByteDance's models as an example: they're closed-source. What benefits does that offer? I can't identify any. Even if the models were open-sourced and fully shared with others, the barrier to entry would remain extremely high—for anyone to effectively use them.

It's extremely challenging for him to implement this solution; moreover, achieving low implementation costs is even more difficult—far from easy. Simply making the code open-source doesn't automatically enable companies to match my deployment costs. There remains significant work to be done. Although the underlying principles are clear, not every company is willing, or has the willingness and capability, to allocate the necessary resources to achieve this goal. This is something I'm quite accustomed to.

It may not be well-suited for this task due to significant resistance. Controlling costs is challenging, compounded by numerous managerial and physical constraints. This actually represents a strength for startups: if a startup is too small, it lacks the resources to handle such complexities; conversely, large corporations often struggle with organizational efficiency.

【00:38:36】

This matter presents significant challenges, making it a key sweet spot for companies of our size. As we grow larger, we may face fewer issues; however, if we become smaller, resources may become insufficient. Regarding open source adoption, I believe we should implement pricing strategies. Currently, we need to recognize that we shouldn 't pressure others—our pricing model doesn't involve excessively high fees; instead, we might charge a fee covering the cost within ten months.

Recovering the investment within ten months alone would render independent third-party implementations unprofitable – something they simply cannot achieve given their costs. Therefore, open source does not affect my revenue. Of course, if I aim for a hundredfold profit, then open source might be relevant... But the video seems to have dropped, sir; it was probably due to a phone call just now. In any case, I believe open source has no impact on our business model.

The premise is that we only earn a sixfold profit and recoup our costs within ten months, which roughly corresponds to a sixfold return. Under this scenario, open-sourcing has little impact on profitability. However, if the goal is to achieve a hundredfold profit, open-sourcing will indeed hinder that objective, as third-party implementations may incur costs twenty times higher than yours. Is this model sustainable in the long term? I believe it can be.

Under this vision, I believe open source can be sustainable in the long term—or at least that's what we intend to do.You can argue that restraint not only helps but also yields long-term benefits. This strategy creates more opportunities ahead of technical challenges and significantly increases our chances of achieving AGI – allowing us to proceed with greater confidence. Consider this: we don't even need to work overtime, as the task isn't particularly difficult. For other companies, however, it could be far more challenging due to excessive complexity in their approaches.

It's actually not difficult at all. Initially, from an external perspective, we might have chosen a seemingly challenging approach—conducting research and tackling the most complex tasks, which appeared like a tough model. Yet by making strategic compromises externally, we remained highly effective and accomplished our objectives effortlessly. Therefore, my conclusion about open source is that it is inherently sustainable.

There is no conflict between open-source and commercial paid models, provided there's no conflict when operating at a sixfold profit margin. While sixfold seems impressive, it's actually not that high. Given the current high efficiency of AI, a reasonable profit margin might be around this level. In the future, it could drop to, say, fourfold or threefold... I believe this is the lowest possible threshold—and still represent substantial profits.

When it comes solely to selling APIs, I don't find that approach particularly appealing. However, you can clarify that there's no conflict—I see none at all. Moreover, I'm not concerned about others deploying our models and competing with us; in fact, we'd even welcome their adoption. We strive to provide comprehensive support to the open-source community to facilitate the deployment of our models.

I'm not concerned that he'll compete with me for this business opportunity, as the market is sufficiently large. My only concern is that he might fail to implement it effectively—perhaps due to flawed details that lead to poor results or high costs. Indeed, there's no conflict here. Last year, when discussing B2B business opportunities, a common question I received was: "If we open-source our consumer-facing product (C-end), will that create conflicts with our existing consumer-oriented product?"

The reason is that we don't have a traffic advantage—Tencent itself has massive user base—and by deploying our open-source model, it attracted all of its end-users, thereby competing with us for our customer base. However, this isn't entirely accurate; there are several underlying reasons. The key question is: Is the open-source model we provide identical to the one we deployed ourselves? Yes, they are exactly the same.

We never claim that using an inferior open-source model and then deploying a better one ourselves creates a conflict – the approaches are identical. This demonstrates there is no inherent contradiction. Throughout last year, we essentially adopted open-source solutions for our consumer-facing services, yet we observed absolutely no conflicts. This exemplifies our commitment to open-source practices. Furthermore, aligning with our company's long-term vision, our ultimate goal should be AGI (Artificial General Intelligence).

While definitions of AI may vary, we can still pursue AGI as our ultimate goal. From a technical perspective, the roadmap for AGI is quite clear: compared to current AI systems, an AI system that can precisely describe a problem with complete context and instructions has already surpassed human capabilities—provided two conditions are met: you provide it with full context and precise instructions.

This definition is extremely difficult to achieve. Take today's meeting as an example: we're surrounded by extensive contextual information—some participants may have decades of experience in that context, something AI simply cannot replicate. While AI can perform better than humans within limited contexts, it still cannot replace human intelligence. The missing element is continuous learning, a capability humans possess through lifelong learning.

When hiring an employee, they typically need two months to familiarize themselves with the company environment and their role. With this period of orientation, they can quickly become proficient and handle various tasks effectively—such as understanding instructions like "Call Xiao Wang" without needing prior context knowledge. In contrast, AI systems lack this two-month learning phase due to the absence of contextual training.

Tell him to call Xiao Wang over, and specify who he is, his position, location, how to locate him, and what precautions to take when contacting him. You must provide the AI with all relevant context. While AI can handle such tasks under these circumstances, it's neither feasible nor practical to supply complete contextual information. Therefore, AI cannot replace human employees. However, if AI possesses continuous learning capabilities—similar to human employees—for example by undergoing two months of training at the company, then it could be effective.

【00:51:02】

However, it can indeed replace humans in various tasks, so we still need to learn how to leverage it effectively. The evolution of AI can be viewed as a series of steps. Last year, we advanced through the CoT (Chain of Thought) approach − realizing that this methodology enables AI to reach higher levels of intelligence by fostering independent thinking, thereby expanding its capabilities. Now, we have moved on to another critical milestone.

This year's progression is centered around Agents, as we've found that the Agent-based approach enables handling a wide range of tasks with broader capabilities and higher intelligence thresholds. Why a step-by-step progression? Because each subsequent stage builds upon the previous one: Agents rely on CoT, which in turn depends on earlier stages—specifically language models—ensuring every step is purposefully designed.

Thus, the evolution of AI and the trajectory of intelligence are predictable. While this year's milestone represents the Agent phase, this stage too will eventually conclude—once it has addressed all potential challenges. However, it cannot replace human employees; its capabilities have reached their limits.

Similar to CoT, once it reaches its upper limit, it surpasses even the most outstanding human performers in solving Olympiad math problems or writing code. Yet it remains stagnant—its technology has not yet reached AGI. Thus, the evolution of AI intelligence follows a discernible trajectory.

Following the development of agents, we identified the key challenge as continuous learning: how to enable models to learn persistently rather than relying on intensive training sessions, allowing them to perform sustained learning akin to human cognition over extended periods. This issue is fundamentally interconnected with task completion and addresses the same core problem. From the perspective of agent development, the next critical bottleneck lies precisely in achieving continuous learning.

The next challenge to address is how to achieve sustained learning—a goal that is clearly identifiable and relatively well-defined. This represents the fundamental obstacle we must overcome, and there is certainly a way to do so, albeit one that requires time. Through continuous learning, we may eventually reach a tipping point: when the model becomes capable of continuously learning, it will be able to perform all tasks achievable by humans.

It can independently develop its own versions, conduct further research, and subsequently create subsequent iterations as well as more advanced artificial intelligence models. Thus, it reaches a singularity where self-iteration becomes possible. However, this "singularity" is not an abrupt event but rather a gradual process—a prolonged evolution rather than a sudden transformation. Yet habitually, we tend to perceive it as a singular breakthrough.

Because long ago, those prophets believed a singularity would occur here, but in reality it is not a singularity; rather, it is a continuous process. Only after this phase is completed do I believe embodied intelligence can truly emerge. This is our hypothesis regarding the likely timeline: first, we must address learning itself; then progress toward an intelligent singularity capable of self-iteration; finally, achieve embodied intelligence. Once attained, embodied intelligence will integrate into the real world, enabling it to perform household chores and provide care for the elderly.

We consider this a fairly ideal roadmap, though opinions may vary and there's no absolute right or wrong. What we value most is that it's the easiest path to follow—one where each step requires minimal effort and eliminates the need for overtime work. However, if the roadmap were reversed—for instance, requiring the implementation of embodied intelligence first—it would become an extremely arduous task. We don't want such a scenario; instead, we aim for a more manageable approach.

If we first address continuous learning, then tackle the self-iterative singularity, and finally develop embodied intelligence, the journey will be much smoother. Later on, we can leverage earlier technologies to aid in developing subsequent ones. Once the singularity is achieved, developing embodied intelligence won't require human intervention—the model will emerge on its own. This precisely answers our long-term goal: what we call AGI.

Let's return to reality. Last year's key reality was that everyone was developing chatbots and competing for consumer traffic. This year, however, the focus has shifted to securing B2B revenue and getting involved in this space—because if you don't participate, you simply won't stand a chance, right?

However, we don't consider this a priority. Within our company, our true focus lies in the AGI roadmap I just outlined and how to achieve breakthroughs in this technology moving forward. Strangely enough, what we're most eager to achieve often eludes us, while what we pay less attention to proves relatively easy to attain. This reflects a strategic advantage: we are fully committed to AGI and dedicated to advancing it.

When developing that application for both the C-end and B-end markets, we didn't need to devote much effort—it actually required minimal resources. From a technical perspective, approaching a relatively lower-level technology from a higher technological vantage point creates a significant advantage. This was precisely what we observed with the C-end market last year: despite not investing substantial effort there and even considering discontinuing user support at one point, we found the users simply couldn't be replaced.

Since they simply couldn't be eliminated, they ultimately remained. However, the revenue from the B2B segment this year appears to show relatively optimistic growth. I believe this figure is likely quite strong compared to our industry peers, in my estimation. Nevertheless, we haven't devoted significant resources to this area—we simply didn't invest much effort in it.

【01:02:45】

I undertake tasks as a natural part of my work. When launching our internet intelligence platform, the development of AGI was an essential milestone—a necessary step on our journey toward AGI. Throughout this process, I integrated these technologies into APIs for public use without adding any extra functionality. Our core focus remains on AI development; this is merely a byproduct.

All I need is a few people to maintain this API—no customer service team, no sales department, nothing at all; users will come on their own. In other words, both end-user (C-side) and enterprise (B-side) applications are by-products of our AGI development journey, serving as intermediate outcomes that don't conflict with our core AGI initiatives.

We didn't develop AGI specifically for consumer (C-end) or enterprise (B-end) applications. Our focus was on AGI itself – and when it happened to produce a viable product, we commercialized it. This sets us apart from other companies that build models solely to serve C-end or B-end users. For us, the original intent was never that; our core mission has always been to advance AGI technology.

To some extent, this represents a strategic dimensional reduction approach. AGI embodies a broader vision that can attract more exceptional talent and offers stronger cohesion. Thus, I leverage my organizational strengths to implement this strategy—a form of dimensional reduction. However, for a commercial company whose core mission is serving end-users effectively, the situation is entirely different.

He possesses other advantages in product development, user service, and traffic volume, but lacks a technological edge. The current favorable situation is that model technology is paramount. You must develop a robust model first; everything else follows naturally, which explains our previous development trajectory. We genuinely chose AGI, and we never intended to serve a large user base. When we suddenly gained popularity during last year's Spring Festival, that wasn't part of our original plan—we hadn't considered such an outcome at all.

Our goal was simply to excel in developing this technology. However, at that time, I realized that our organization actually possessed distinct advantages in terms of talent and organizational structure compared to fully commercialized, product-oriented entities—a truly remarkable finding.At that time, everyone was fiercely competing for the C-end market, only to see it ultimately captured by someone who didn't even try. This indeed confirms the logic I mentioned earlier: there truly exists a competitive edge in terms of both talent and organizational capabilities.

This talent advantage doesn't mean my team members are inherently smarter than theirs; rather, it lies in how we organize these talents, motivate them, and foster collaboration. This approach offers distinct advantages. Simply gathering intelligent individuals doesn't guarantee natural cooperation or passionate pursuit of a common goal—thus requiring a clear vision. Our previous experience has shown me that the AGI vision is exceptionally powerful. Now, that's the key issue.

The next question is: what exactly are the importance of our core interests? As I mentioned earlier, we need to exercise great restraint in many aspects. But what are our core interests? In fact, there is only one: maintaining team stability is our paramount core interest – indeed, perhaps the sole one. As long as we can preserve team stability, we will surely achieve success with AGI; that's simply it.

As long as everyone remains committed, we can continue our work with virtually no risk. The only uncertainties lie in timing – setbacks may occur sooner or later; but if all parties stay engaged despite challenges, we can proceed again. Funding and resources are not issues; other essential elements are readily available. For us, there is only one core priority that cannot be compromised: maintaining team stability.

This is also a significant challenge we face, or rather, I consider it the greatest risk. Of course, this risk has been substantially mitigated by our recent funding round. Since everyone has received relatively substantial option packages, the overall impact is limited. Regarding team stability, as long as key and most experienced employees remain in place, others are unlikely to leave; even if their options or compensation packages are smaller, they are still less inclined to depart.

The primary motivation isn't solely financial; everyone aims to work within an environment conducive to developing AGI. Therefore, this field remains highly attractive to talent. Historically, our talent mobility has always been relatively low compared to industry peers. Yet this remains our greatest—and indeed the only—challenge. The rest are merely time considerations; at most, they could delay us by six months or a year, but certainly not prevent us from achieving success.

There is certainly no shortage of funds or resources—in fact, all these are readily available. Therefore, much of our current work focuses on maintaining team stability. Beyond this, I believe we can do without everything else and exercise restraint in other aspects. We have always been very restrained, unwilling to compete with any major or smaller internet company. My goal is to empower them or assist everyone in achieving this objective.

This also constitutes part of the commercial significance we mentioned earlier, provided that everyone complies with the requirements.

【01:15:12】

Under these circumstances, we are more than willing to assist anyone—even our competitors such as Alibaba, ZhiPu, and Moon Dark Side—in helping them improve their work. We gain nothing from this; after all, we operate on an open-source basis. Open source inherently does not define boundaries clearly enough; what matters is whether you can reproduce the results. If you cannot reproduce them, we'll guide you on how to do so. This is simply part of the open-source spirit—it shouldn' t be affected just because you're a competitor.

Of course, if we're partners, I'll take on additional responsibilities. However, there's no conflict regarding major interests. When working with external parties, our stance is clear: we focus solely on the core aspects of AGI—such as GPT, CoT, Agents, and so on.

The field of AI is vast, and there are many areas we believe don't align with our core focus—such as 3D modeling and video generation—which I think have little relevance to the main direction of intelligence development, so we won't pursue them. Similarly, world models currently don't significantly contribute to advancing the boundaries of intelligence, so we'll not develop them either. However, if others undertake such projects, we'd be happy to assist. Time availability is one factor, but there's no conflict of interest.

We also hope these AI technologies can be deployed across various production environments to enhance societal productivity and assist industries in improving operational efficiency. We are highly committed to this endeavor. Whether we have the time, personnel resources, or whether our colleagues are interested is another matter entirely; however, there are no conflicts of interest—we are solely focused on achieving this goal.

We believe this does not conflict with commercial interests at all; the benefits I should receive have not diminished. Maintaining this stance has not resulted in any reduction of our gains—not because we adopted open-source practices, nor due to our goodwill or assistance to others. For instance, our user base among end-users remains substantial and stable even last year.

I'm quite optimistic about our B2B business this year. Our good intentions haven't compromised our commercial interests at all—in fact, they may even have brought additional benefits. This might seem counterintuitive, but it's true. Conversely, if we consider the opposite scenario: would going against this approach necessarily yield better results? No. The key question is: how can we understand that there's no direct correlation between the world model and improvements in AI intelligence capabilities?

This is our assessment at the current stage. We recognize that this represents our own AI roadmap, though not the only one available. In our view, the top priority now is to excel in AI training. Achieving excellence in AI training does not require a global model or even multimodal approaches—by narrowing the scope of AI training and eliminating multimodality, certain tasks may remain unachievable without compromising the algorithm's validity.

Multimodal approaches ultimately need to be implemented. The current priority is training; the next step involves addressing continuous learning, followed by enabling autonomous question-asking capabilities. However, this roadmap does not include world modeling or video generation. When video generation first emerged, it became a hot topic—almost considered essential; failing to implement it would render one ineligible as an AI company.

So I find it quite puzzling—upon closer reflection, this actually has nothing to do with the Smart Roadmap.In fact, you'll notice that when video generation first emerged with Sora, it was adopted by everyone—from large corporations to small startups. However, most smaller companies eventually discontinued it. This has nothing to do with the limits of artificial intelligence; rather, it's purely a commercial success—a viable business model.

We don't pursue a project simply because it's commercially viable; we only undertake it when it aligns with our strategic roadmap for artificial intelligence development. Take video generation as an example—it's relatively straightforward to define. In contrast, the concept of a "world model" is less clear-cut, as many approaches can be categorized as such. Based on our assessment, developing world models and achieving true AI sophistication remain not the most critical priorities at this stage.

The most critical aspect lies in AI training and how to ensure continuous learning post-training. This is our company's perspective, though each organization may have its own view. As I mentioned earlier, the foremost challenge for us is personnel stability. From another angle, what exactly do we lack? What sets us apart from U.S. companies? The difference is minimal: resources. We don't have as many resources; our pool of talent remains relatively limited.

Our current computing capacity is approximately 20,000 H-equivalent units, most of which have just arrived within the past month or two, with many more machines yet to arrive. While our total computing power was relatively limited last year, we are now expanding it aggressively this year. We currently operate at around 20,000 H-equivalent units and plan to procure a significant number of additional GPUs in the coming months—primarily from NVIDIA. The more GPUs we need, the better.

Within our financial capacity, it's undoubtedly true that more cards are always better. Our current strategy is to purchase as many cards as possible at a reasonable price—exactly how many we can afford after using this funding round. The spending pace isn't predetermined; we'll buy whatever is available as long as prices remain competitive. In fact, I'd consider that a positive outcome if we spend the entire amount within six months.

If I could spend all the money within six months, that would be truly blissful and ideal. In reality, spending such a large sum is no easy task: you can't obtain enough cards, they're hard to come by, and their prices are quite high—not necessarily exorbitant, but certainly reasonable.

【01:26:41】

If I can spend this money within six months, that would be ideal. After all, converting the funds into an NVIDIA card is definitely better than keeping them in a bank account – my current interest rate there is probably only around two percent. However, purchasing an NVIDIA card involves a ten-month upfront cost. Therefore, it's best to buy only what you can afford. Once you've acquired the card, you'll have ample flexibility; through providing services or other means, you'll consistently generate cash flow.

With sufficient cash flow, I can live comfortably without holding large amounts of money in my account. Therefore, our only concern is whether we can obtain enough cards. If converting all funds into cards were feasible, we would undoubtedly do so without hesitation and are even willing to pay a premium for this benefit—it's simply too cost-effective. Even after paying the premium, however, achieving this goal remains challenging.

Objectively speaking, if I can spend two billion this year, it would indicate that our procurement department has achieved outstanding performance. The main gap between us and the United States lies in resources, while the disparity in personnel is minimal—there is virtually no difference, as we are essentially the same team of people, possibly from China. When these Chinese employees travel abroad, some remain domestically, others stay overseas, and still others go abroad; it's not necessarily the more capable individuals who leave, but rather a relatively random distribution.

The most talented individuals likely do not travel abroad in full measure, but rather remain domestically to a greater extent. There is no shortage of domestic talent, and given our large population base, we continuously attract new talent each year. Talent itself is not the bottleneck; resources are the primary constraint. Resources fundamentally impact talent development—due to limited computing power, opportunities for experimentation are limited, resulting in an overall talent gap compared to the United States. This talent disparity essentially stems from a deficiency in computing capabilities.

With the largest models available today, we simply cannot afford to train them. Even if we spent all five hundred billion yuan, we still wouldn't be able to do so. Even if we could accumulate the resources, we wouldn't have the means to utilize them. The current largest model requires approximately 800 billion activations; domestically, we are still at a scale of several dozen billion activations, and even the largest domestic model may only require several dozen billion activations—a difference of an order of magnitude. To train a model of the same size as an AI system, we would need around 50,000 GB300 GPUs or Huawei 950 GPUs, totaling two hundred thousand cards.

This is merely training; research has not yet been considered. Therefore, the biggest gap between us and the United States lies in resources. Our current resources, as well as those available within this year and the coming months—including substantial additional resources—we only have enough to conduct further experiments at a scale of ten billion activations. At that scale, there are still numerous experiments to be carried out and many aspects that need clarification.

We are still far from developing a model capable of training 800 billion parameters, as there is ample time and sufficient computing resources available. In my view, the fundamental difference between us and the United States lies in resource availability. All observed disparities—including differences in talent pool, model capabilities, and application scope—can be attributed to variations in computational power resources. On one hand, domestic hardware components for such computations are unavailable; on the other hand, our capital investment remains significantly lower than that of the United States.

We have significantly reduced our capital investment, and the proportion of talent salaries in this context is very low. Although they offer salaries as high as $100 million, when calculated, talent compensation still constitutes a minimal share; the majority comes from computing power. This issue remains largely unsolvable at present, as Huawei's production capacity is also limited.Because if I need to train the 800B model, I would require 200,000 of Huawei's latest cards. This is merely for training purposes and does not yet include considerations for research applications.

Therefore, we simply do not consider competing with the United States at such a large scale. Our current shortcomings lie in our ability to operate at a scale that is both feasible and affordable for training—specifically, achieving activation scales of several dozen billion users first. Only when we have more resources in the next phase can we expand to scales of 150 billion, 156 billion, or even 250 billion activations. Consequently, there remains a gap between us and the United States that appears difficult to bridge at present.

You can certainly train large models, but you cannot conduct thorough research beforehand. Another key concern is: what will ultimately define the competitive gap among all large models? In my view, this gap may not be significant overall. The final differences will likely manifest in three areas: cost, time investment, and user experience.

Beyond that, there may not be any significant difference. Cost is relatively straightforward to understand: when providing the same service with identical quality, what price point can you charge? Regarding BYD's batteries—for the same technical specifications—whether competitors can offer them at comparable prices is quite challenging; it's certainly not easy to achieve, and this undoubtedly creates a barrier. Therefore, cost is undoubtedly a key differentiator, and I believe it ranks as the primary factor distinguishing one product from another.

The second factor is time—when you can actually deliver the product. Whether you launch it months earlier or later makes a significant difference. The third aspect is user experience, which may vary somewhat. While there will still be user retention challenges and barriers to adoption, these may not be fundamental issues. The core considerations remain the first two factors: cost and time. It also depends on whether you develop the product first or later, and whether you implement the same features more quickly or slowly. Once the long-term commercialization strategy and product portfolio are further refined, how should pricing be determined?

We believe that the most worthwhile and effort-intensive endeavor at present remains developing AGI. The current priority is to push AGI forward further—specifically, to elevate its fundamental capabilities. At this stage, this approach proves more cost-effective than developing additional product lines or exploring various commercialization pathways; indeed, it represents a path with greater returns. We anticipate this will hold true for both the foreseeable future and any previous period. In other words, if we proceed...

【01:39:46】

We spent considerable time reflecting: does enriching our product actually serve any purpose? What commercialization strategy was we even discussing half a year ago? It inevitably meant I'd need to run advertisements, develop e-commerce features, integrate e-commerce elements into the product, and achieve deep integration with local lifestyle services. None of these approaches proved viable, given how rapidly things are changing. At this stage of our product's development, dedicating significant effort to commercialization considerations or mapping out specific pathways would only result in a very short product lifecycle.

I don't think this is the right time yet. This is our assessment, or at least previous experiences all support it. Over the past three years, whenever you've approached me about the commercialization strategy or product roadmap, it's been a waste of time—because we can neither predict nor foresee the future. What we can anticipate is very limited. As for my perspective on timing: should we take a break? Have some food first? Or continue the discussion? If there's no disagreement, I'll proceed.

We are conducting world-leading research on AGI and pursuing commercialization at the appropriate time. I believe we have consistently been working toward commercialization—though not with it as our primary objective. Our focus remains on advancing AGI, yet our persistent efforts in commercialization have enabled us to acquire end-user customers (C-end) and generate business revenue from enterprise clients (B-end). Historical experience has proven this strategy successful. Furthermore, I believe the transition to full commercialization remains a distant prospect.

The most significant aspect lies in the extension of technology and the development of next-generation solutions designed to address current challenges more effectively. In my view, these benefits are already evident in the foreseeable future; focusing solely on products at this stage would be premature. This is precisely why we exercise restraint.

We hope these business opportunities can be pursued by others. The utilization of AI should be a collaborative effort involving the entire society and all our partners, with shared benefits rather than monopolization – that's simply not feasible. Moreover, we don't have sufficient resources or organizational capacity to handle this task alone. As for our partners, they are carefully selected through our rigorous screening process.

Regarding the specific proposal for cooperation, I believe that our interests are largely aligned—specifically with those that best serve our own interests, are least hostile toward us, or most encourage our success. Not everyone wishes for our success, as we have indeed harmed the interests of many others. This applies to the processes and decision-making mechanisms governing major corporate strategies, technological initiatives, and business decisions.

Our company operates fundamentally on a culture of consensus. I don't make all decisions alone; rather, I strive to build shared understanding. My authority and influence within the organization are rooted in this collective agreement. For instance, before taking any action, I always first assess what the team agrees on and whether everyone supports it.

While there may be some guidance or tendencies involved, their impact is limited—indeed, very limited—and relies entirely on consensus. This decision-making mechanism fundamentally serves to foster consensus; it's not about me pushing something forward; only when consensus exists can I proceed with implementation. Currently, my primary focus is entirely on DeepSeek. Let's take a five-minute break—I' ll get back to writing shortly.

Everyone, please turn on your microphones and share your feedback. If there are no issues here, we can proceed; otherwise, let's take a five-minute break. The significant differences in final model outcomes among various manufacturers likely stem from comprehensive factors. To compare model performance effectively, comparisons must be made at the same cost level – that's what makes it meaningful. After all, when comparing two vehicles, you're comparing models within the same price range. The difference between well-developed and poorly developed models shouldn' t lie in specific components but should be reflected overall.

Anthropic has now surpassed OpenAI—but is this a long-term trend? I don't believe so; this is definitely an extreme case. OpenAI and Google are likely to continue alternating in terms of performance growth going forward.In reality, Anthropic's advantages on Code Agent aren't as significant as they might seem—and it doesn't even dominate OpenAI outright.

About half of our employees generally believe that OpenAI is superior. While Anthropic does enjoy a first-mover advantage, this edge is likely to fade quickly and cannot be sustained long-term. All three companies are highly competitive, with Anthropic demonstrating the highest operational efficiency—incursing the lowest costs and expenditures.

While dominating the division of labor in global AI, Chinese companies are highly likely to play the role of having the largest production capacity. Logically speaking, China possesses the greatest production capacity—including in chips—and may also have the most electricity supply, making it highly probable that China will become one of the three major players in AI. Chinese manufacturers will strive to produce these products at the lowest cost while ensuring optimal performance, especially since many foreign products manufactured in China now differ little from those produced in the United States.

The same may hold true for AI in the future, but AI developed by China might be more affordable. This affordability could be systemic, similar to how services provided by China are often cheaper in other industries. When I undertake tasks, my habitual approach is: What will yield the greatest benefit at this moment? If I believe developing a product will maximize benefits now, I will focus on that; if I think implementing AGI will bring the greatest returns first, then I will prioritize AGI development.

Clearly, I believe developing products at this stage is not the most profitable approach. Regarding compatibility issues with domestic graphics cards, there actually exists a historic opportunity for them. Previously, the main challenge was the weak ecosystem: users could purchase these cards but couldn't utilize them effectively due to the lack of NVIDIA's robust ecosystem support. This highlights NVIDIA's significant competitive advantage. However, this situation is now changing.

NVIDIA's CUDA competitive advantage is rapidly eroding, likely due to three factors. First, the advent of AI has significantly simplified ecosystem development—AI enables automatic code generation, making ecosystem building far easier than before.

【01:56:36】

I can leverage AI to build this ecosystem, replicating exactly NVIDIA's framework. First, it's powered by AI; second, it incorporates cutting-edge technologies. For instance, we've developed TileLang—a high-level programming language that enables rapid implementation of all NVIDIA's ecosystem components when writing CUDA operators. When combined with AI, this approach presents virtually no challenges.

However, this technical approach has not yet been fully realized. Nevertheless, it appears to face no significant obstacles. Additionally, since CUDA originated from graphics cards, its design and configuration share a direct lineage with those of gaming GPUs, ensuring full compatibility. Previously, AI computing was a relatively niche field compared to the broader graphics card market, making this approach entirely reasonable.

However, the market for computing cards has now surpassed that of gaming cards, leaving no reason to maintain coupling between the two. The current trend is that such coupling will cease in the future. Consequently, dedicated chips—whether from Huawei or NVIDIA itself—will become exclusively dedicated chips, rather than the previous hybrid solutions. In this context, the role of NVIDIA's ecosystem has significantly diminished. For a dedicated chip, it has little to do with CUDA and is not bound to it.

In other words, the chip's design inherently incorporated considerations for building an ecosystem. While somewhat complex, there is indeed a historic opportunity for domestic AI chips to replace foreign alternatives. We believe that within the next year, one thing will be proven: the ecosystem of domestic chips poses no significant challenges at all.

Initially, it was perceived as problematic—considered impractical or difficult to use—but I believe we can change this perception within a year, supported by concrete evidence. Domestic AI chips have no issues with hardware or ecosystem; the only challenge lies in insufficient production capacity. When it comes to card compatibility, there are no obstacles—the NVIDIA cannot hinder progress.

In a normal commercial environment, if I could purchase an NVIDIA graphics card, domestic alternatives would be quite challenging to develop; however, when NVIDIA cards are unavailable, everyone is forced to turn to domestically produced chips. Under these circumstances, there are no obstacles to the compatibility of domestic graphics cards. Domestic cards can establish an ecosystem comparable to, or even superior to, that of NVIDIA—I believe there are no barriers, though it will take time. Currently, our primary collaboration is with Huawei.

Huawei will adapt its own solutions, but we will actively participate in this ecosystem and engage deeply within Huawei's framework. The core issue for Huawei remains insufficient production capacity. For instance, Huawei has provided us with a production capacity of approximately 16,000 cards, whereas major internet companies may produce hundreds of thousands of cards—our output is only around 10,000 cards. I believe this ratio is quite significant, but this already represents the maximum capacity Huawei can offer.

Therefore, we cannot expect to train the larger model mentioned later on Huawei, or to train a model with hundreds of billions of parameters activated—this has been discussed this year. However, there might be opportunities next year or even the year after. Regarding Huawei card adaptation, our primary focus is to improve its high-level language compiler and refine TileLang. Once TileLang is optimized, the issues will likely be resolved effortlessly.

This matter is quite complex, but we're working on it now; once completed, the explanation will become much clearer. As you can understand, during V3 training, NVIDIA GPUs were still used, but the NVIDIA ecosystem was no longer employed.

The V3 uses NVIDIA's graphics cards but does not leverage NVIDIA's ecosystem. Instead, we first developed a high-level compiler called TileLang and then built all other functionalities around the TileLang ecosystem, thereby achieving near-complete independence from NVIDIA's ecosystem. Simply replicating this entire process on Huawei cards completes the solution.

I believe this represents a historic mission: to completely change the prevailing perception that China's gaming card ecosystem is weak.Now we have basically all the necessary conditions—only time remains. I believe within a year, many players will be involved or recognize that this issue has been resolved; the remaining challenge is capacity expansion. I'm quite optimistic about domestic computing power development. In this regard, NVIDIA is essentially digging its own grave.

Huawei's hyper nodes, specifically the Huawei 950 hyper nodes, can fully replace NVIDIA's GB200 and GB300 in terms of performance and price. The cost is certainly higher, but only marginally so. Whether it's 50%,100%, or even 200% more expensive doesn't matter much. For instance, a price increase of 100% would already qualify them as viable substitutes in terms of cost.

They are also interchangeable in terms of tasks; all tasks that a GB300 can perform, a Huawei hypernode can do as well, with identical latency performance. The only cost is that four Huawei cards are equivalent to one NVIDIA card, and they lag behind by two years in performance. The equivalence of four cards to one is understandable. The "two-year lag" means that four Huawei Huawei 950 cards can match the performance of one GB300 card, while the "two-year lag" refers to a temporal disadvantage of two years.

The Huawei 950 supernode will be available in Q3 or Q4 this year, while the NVIDIA GB200 was released in Q3 two years ago—there's a two-year gap. NVIDIA may have already introduced its next-generation model by Q3 this year. Therefore, regarding the gap between us and the U.S. in chip technology, I believe there won't be any further disparity at the ecosystem level, but the gap in chip technology remains four times larger with an additional two-year delay. Another question is whether we will pursue vertical integration upstream. I hope not.

So the question arises: should we pursue vertical integration across upstream applications? We prefer not to do this ourselves; we'd like others to handle it. I don't want to absorb everything—I'd rather focus on just one area, specifically the one where we excel or what we consider our core strength, and then deliver it directly to users.

【02:08:18】

Regarding the direct relevance, I believe that for many of our industry partners, their focus often lies more on what others consider significant than on how we interpret it ourselves. Will we establish large-scale industrial clusters in the future? I am convinced that building such clusters is essential—we have been actively pursuing this initiative, and all our clusters have been self-developed. However, whether to develop our own chips will ultimately depend on the potential returns, which are crucial.

Tesla is focusing on this very aspect. If you operate a power plant, you don't necessarily need to manufacture your own generators, right? The power generation equipment can be sourced from third parties—as long as the price is reasonable, why would you produce it yourself? Therefore, I hope we don't have to develop our own chips; instead, I want to purchase them at reasonable prices, thereby eliminating the need for in-house chip development.

I believe the scenario is likely as follows: while NVIDIA's chip profits may not be sustainable in the long term, we aim to focus solely on this area. Artificial intelligence is a vast field that doesn't require me to concentrate exclusively on one aspect. With proper focus, and given the substantial business potential here, the AI era will give rise to numerous trillion-dollar companies – and we are poised to be one of them.

We have always focused on a small part of this endeavor; being one of the participants, we face no significant challenges and don't need me. I'm not overly ambitious about establishing a prominent position. I believe it's likely quite challenging to change course, and if you truly want to pursue something different, you'll develop even better ideas—this reflects our approach.

For instance, at least in the B2B and B2C sectors, we've observed that companies genuinely committed to building complete B2C ecosystems often achieve superior results, whereas those focusing solely on B2B closed-loop solutions may not perform as well. The more ambitious your goals are, the more challenges arise—particularly regarding B2B operations. The potential of B2B business ultimately depends on demand; given current advancements in AGI and AI technologies, B2B demand remains inherently limited.

It will grow rapidly, but not indefinitely; ultimately, its growth is driven by demand rather than computing power. Given current technological conditions, revenue levels ultimately depend on market demand. From your perspective, I could recoup my costs in ten months – that's precisely why I'd invest upfront if there's a viable opportunity, especially when resources are limited. Indeed, demand is likely to continue expanding as technology advances steadily. We've been consistently pursuing a multimodal strategy.

This feature is crucial for both the product itself and consumer-facing applications. However, regarding the smart capabilities' limitations, it serves as a component rather than the core functionality. Nevertheless, as a component, we will undoubtedly implement multimodal support—and we are already doing so. We plan to develop relevant models, ensuring that versions like V4 and subsequent iterations will natively support multimodal functionality.

However, when it comes to multimodality and intelligence, we view them as components rather than intelligence itself; their core functionality aligns with search. Search is a component, so is multimodality—we fundamentally understand both as distinct components. Regarding scaling, we firmly believe that greater scale inevitably leads to better performance and unlocks more capabilities.

The real obstacle to scaling is computing power—not a lack of willingness to scale, but insufficient computing resources. We haven't yet identified the true limit of this capacity. Training such large models isn't because we believe they're sufficient; it's simply because we have the necessary resources available.

I calculated this based on my available resources—the size of models I can accept and train—that's how it was determined; it doesn't mean this model alone is sufficient. Currently, the benefits of models remain quite evident. We haven't yet encountered the wall of scaling; that's still a long way off for us. When Silicon Valley talks about reaching the limit of scaling, that's relative to Silicon Valley itself; for people in China, we're still far from that point—we haven't even scaled to that level yet.

This scaling encompasses data scaling, model size scaling, and training costs. We are still far from exploring the full potential of this scaling due to limited computational resources, but we will spare no effort to push its limits.After completing our funding round, we'll have more computing resources and can potentially train larger models.

The sooner you purchase it within the specified timeframe, the better; ideally, if you can devote all your efforts over six months, that would be ideal—but in reality, this is unattainable. In my view, the core capability of a next-generation model must include continuous learning ability to truly qualify as such. Until then, our focus should be on reducing costs while improving performance and accelerating speed. However, achieving a major breakthrough requires fundamentally possessing continuous learning capabilities.

Another question arises: why do we seem particularly concerned about a model's computational efficiency? The reason is that I've observed not everyone cares much about this aspect. For a commercial company, there's little incentive to pursue high model efficiency—after all, a slightly better efficiency doesn't provide significant competitive advantage.

So for several startups, they don't explicitly state that low cost is their goal, because it doesn't align with their interests. If costs are too low, how can you make any profit? Conversely, achieving low costs would mean losing revenue altogether. On the contrary, this is part of our vision. Our team members genuinely care about cost efficiency, as they are ordinary individuals who understand that all business operations involve expenses.

He demonstrates this empathy: if others have to pay for a product, they are more likely to accept it at a lower price. Therefore, many of our colleagues still expect us to reduce costs further. However, from a business perspective, this isn't how we think about it.

【02:21:31】

From a cost or commercial perspective, this is not the top priority. For both startups and large companies, the service cost is simply not high. However, I believe we want it to be lightweight and affordable—especially given the current shortage of computing power in China—so that it can be used on domestically produced GPUs. I think low cost should be the primary outcome.

Our model architecture has consistently evolved toward lower costs, which aligns with our strategic vision. We also employ various algorithmic approaches that further reduce expenses. Another key factor is that lower costs enable us to train larger models and support more extensive architectures. Given limited computing resources, higher computational efficiency allows us to accommodate even larger models.

For large corporations, this perspective may not apply. While resources can be allocated to address such challenges, we prioritize cost efficiency above all else. The value of data-driven models lies in their extensive scope—data itself constitutes nearly half of a model's value. To illustrate: if AI is projected to account for 20% of GDP, attempting to allocate merely 5% within that framework would simply be unfeasible.

I'm certain I'd be defeated by another competitor—if they claim I only need to achieve one percent of the total, I'd definitely lose. However, if my goal is to account for five percent of both AI development and global GDP, this calculation holds theoretical validity. Take OpenAI as an example: their calculations seem plausible and theoretically sound. Yet they face a critical limitation—they could be defeated by anyone willing to contribute merely one percent.

When one person claims, "I perform so well that I only need to contribute 1% of global GDP," they will be outperformed. If another person then states, "I only need 0.1%," the first speaker will again be defeated. From a macro perspective, regardless of the specific percentage involved, there is no difference between them—they all yield identical results. Those who contribute more are ultimately outperformed by those who contribute less.

You don't even need to secure substantial resources—if your vision demands extensive investment, you'll be outcompeted by those with limited vision. In reality, no one gains tangible financial benefits; it's merely a shared vision. Those with ambitious visions often face early setbacks and greater challenges. That's simply how the world works. OpenAI initially believed it could dominate this field, but in practice, it encountered numerous competitors. Every challenge it faced made its path far from easy.

The challenges the United States has faced in the past may still arise for it in the future with China, because Chinese people are willing to accept lower compensation in exchange for providing this service. In China, there will also be individuals willing to accept slightly less. However, a balance must ultimately be struck: if you receive too little, the company's business model becomes unsustainable and it cannot survive; if you take too much, you will be outcompeted by those who accept lower pay.

Therefore, for us, our goal is not to set prices that maximize profits or returns, but rather to achieve a reasonable profit margin. This is my explanation. I firmly believe in this approach; I'm not trying to justify it—there's no need to do so. This is simply how I operate, and there must be a reason behind it. The rationale may not be conventional, but I believe it's precisely what companies should strive for.

Our company's management operates along two distinct axes: one top-down and one bottom-up. The bottom-up approach allows each individual to pursue their own initiatives without supervision or KPIs. The top-down approach refers to formal processes where collective action requires coordinated efforts across the entire organization. For instance, launching version V4 necessitates clear division of responsibilities with each person contributing a specific portion. This dual framework constitutes our formal management structure.

Generally, we aim for this initiative not to consume more than half of employees' total time; indeed, it should not exceed that limit. The remaining half of their time remains unallocated, allowing them to pursue whatever they wish. This constitutes a research scope that enables employees to explore independently based on what they consider important, without any pre-established requirements. As long as the company provides adequate support—either in terms of computational resources or by requiring minimal effort from employees themselves—there is no need for coordination at all.

This is how we are organized today. Some people view our approach as top-down, while others see it as bottom-up; I believe both perspectives are valid. My rule is that formal meetings should not exceed half of the total workload. We generally avoid overtime work as well. There are two reasons for this: first, conducting research requires a relatively relaxed environment; excessive pressure would hinder effective research.

Since this requires genuine personal interest and regular reflection on these issues, exploration can only occur in a relatively relaxed environment – a necessity stemming from the demands of cultural research. Secondly, we maintain extreme focus; such intense focus means we undertake very few tasks. Consequently, I have little to do and thus don't need to work overtime. This approach is consistent with our previous principle of self-restraint: through restraint, I often refrain from taking on additional tasks altogether.

As a result, I have fewer tasks to handle, and each team member is assigned less work. You may notice that many of our products are still incomplete, yet we haven't attempted to address these shortcomings – this reflects part of our culture. Okay, given the numerous issues, I've reviewed most of them. If you have any questions, please feel free to ask. Investors are welcome to speak freely during the discussion. One important reminder: Liang Wenfeng has shared some sensitive information; please refrain from disclosing any figures or details, including those related to card volumes.

Also, do not record your screen for external sharing. Thank you all very much. If you have any questions, feel free to join the discussion.

【02:34:59】

Mr.Yang, could you provide a timeline outlining when sustained learning leads to breakthroughs? Specifically, what additional architectural and algorithmic innovations are required to achieve sustained learning, along with other key elements? Given the limited number of researchers, this field demands extensive investigation—indeed, it's a topic of global interest. For investors, the term most frequently encountered is "agent"; however, for researchers like us, the focus lies more on learning mechanisms and how to address challenges related to learning.

In reality, learning may not be a single technology but rather a fundamental problem. There are numerous approaches to solving this problem—it isn't confined to one specific technique or entity, but encompasses multiple components. Similarly, AGI consists of various elements: it requires models as well as numerous other components. Essentially, it represents both an engineering challenge and an algorithmic issue. While highly specialized, there are extensive methodologies and substantial research efforts dedicated to it. Thank you, Mr. Liang, for this opportunity today.

First and foremost, I would like to express my sincere gratitude for your opening remarks, which deeply moved me and provided us with valuable insights. You mentioned that this team operates with the utmost goodwill, aiming to contribute modestly yet meaningfully to advancing society and the development of human intelligence within this industry, driven by a strong sense of mission and vision. This aligns closely with the corporate culture of our company, which embodies the principle of "cultivating oneself to benefit others."

I have deeply understood why you chose to lead your team in pursuing open-source initiatives. I can vividly envision our endeavor as creating an ecosystem akin to a banyan tree—a haven for birds that benefits all without claiming dominance. Such an approach ensures widespread acceptance, fosters coexistence among all elements, and ultimately leads to universal adoption. Through this investment, we aim to demonstrate our recognition, support, and respect for this mission and vision. Simultaneously, we aspire to contribute meaningful efforts in areas where we excel within this industry's future development.

I would also like to continue consulting and discussing this matter with you. For instance, in the future co-development of the ecosystem, given that open source has now been adopted, how many partners, talents, and teams in this industry can effectively replicate our current open-source models and achievements? As we move forward with further ecosystem development, in what areas do you think more high-quality professionals are needed to integrate into our models and reproduce them? Or is there currently a relative scarcity of GPU computing power?

Will the future of models follow a matrix-based approach? For instance, as we continuously refine the foundational models of large-scale AI systems, partners and teams across various industries can develop specialized industry-specific models or application-specific models within this framework. What is the current state of development in this area? Over the next two to three years, what kind of ecosystem do you envision will emerge? This is my first question for you. Secondly, you have just shared numerous insights with many partners regarding AI hardware developments.

Like major global AI companies, each may invest hundreds of billions of dollars individually. Currently, China appears to have certain shortcomings in hardware computing power. How long do you think it will take to address these issues and support the development of our AI, ensuring that limitations in computing power and hardware do not hinder the advancement of AGI? Do you believe this is a mission that Chinese researchers must accomplish in the future—something we are definitely capable of achieving, albeit with challenges regarding time and funding? However, the outcome may involve both aspects.

On one hand, advancements in models will enhance their intelligence, gradually reducing the hardware and computational resources required for single tasks or specific intelligent components—making high-computational-power operations less necessary as models become more efficient at processing data. I'm not sure if my understanding is correct. On the other hand, technological progress in hardware will improve computational efficiency. Could these two trends be moving in opposite directions?

At this stage, whether using the current 960 or H200 processors to deploy computing power investments worth hundreds of billions of dollars, as you mentioned with a three-year amortization period – what is the actual lifecycle you estimate for these systems? Do you think the technological iteration cycle spans four to five years? Or put more bluntly: if a data center were built with today's GPUs to form a 10,000-card cluster, after three years it might essentially operate on relatively outdated computing capabilities?

Could it happen that at this stage the computing resources are still under construction and insufficient, only to become relatively less high-quality with excess capacity three years later? I'm not sure if such a scenario might occur. I'd like to ask you about these two questions—thank you. The first concerns the ecosystem: we believe every company currently faces a talent shortage, but I think this shortage will be temporary.

In the early stages of development across all industries, there is always a shortage of talent. This was particularly evident in website development: when it first emerged, very few professionals worked in this field, and talent was scarce. Later, when the internet shifted focus to server-side development, the same issue persisted. However, such talent shortages typically resolve themselves within two to three years as a large pool of professionals emerges. The AI talent shortage is also temporary, and we have already witnessed a significant alleviation of this issue.

There is no genuine shortage of AI professionals; every company can rapidly cultivate talent, and the training process is swift. Consequently, the entire AI industry—whether in terms of ecosystem development, model companies, or other aspects—does not face a talent deficit. Any perceived shortage is merely temporary; historically, there has never been a prolonged scarcity of specific types of professionals. I recall that over a decade ago, there was concern about a pilot shortage due to the lengthy training period required, but this issue was quickly resolved. Therefore, there is no need for undue concern regarding talent shortages.

Moreover, there are currently too many companies in China engaged in model development—indeed, far too many. In the United States, there might be only three such firms, whereas China produces an excessive number of basic models. Ultimately, there will inevitably be no need for so many people to develop these fundamental models; this trend will surely converge. Consequently, resources remain relatively scattered and, to some extent, wasteful.

【02:44:00】

Every company has to do the same thing, but in the United States, only three companies need to do it, so resources are concentrated solely on these three. China's resources are highly dispersed, and each company receives even fewer resources. I believe this trend will inevitably converge—but it requires a process, and ultimately it will happen. There's no need for so many companies, because currently, many may perceive the profit margin from this endeavor as extremely high, hence they insist on doing it themselves. However, once they realize the profit potential might not be as high, they may abandon the initiative.

There definitely aren't such high profit margins nowadays; I simply don't believe they exist, as this contradicts objective laws. This indicates we're at a stage where extremely high profit rates would inevitably violate these principles—we should aim for reasonable profits instead. This reflects the current state of the industry, and I'm convinced it will eventually converge. Regarding the development of large-scale models, it's unacceptable for any single entity to monopolize them or claim, "I want to capture all the profits."

If each company only charges a reasonable profit, then there would actually be no need for so many players to develop large-scale models. In the end, China would have only three or four competitors, and competition would be quite intense—the prices would certainly be sufficient for a price war. For large-scale models, perhaps it's enough to have just two major companies or two smaller ones. As for the ecosystem, I don't have many specific ideas. We hope to support more participants, but we don't have the necessary resources to do so.

We have this intention and there are no conflicts of interest, but whether we actually take action is another matter. At least there are no conflicts of interest here; our goal is mutual benefit and win-win cooperation. First, I absolutely do not believe that large model companies can capture the majority of profits—this is impossible, given that the gap between these companies is not that significant. The only distinguishing factors are time and cost. Therefore, none of them can achieve excessive profits; I don't think that's likely.

Those who excel at cost control earn more, while those who perform poorly earn less—that's simply the rule. Has anyone provided an answer? Has the first question been addressed? I'm confident many will now understand that the future holds great potential for data-driven applications through this iterative cycle. Can we hear that now? Thank you. Indeed, your response is insightful, and I deeply appreciate your profound understanding and respect for your strategic positioning within the industry ecosystem.

Take data as an example: currently, model companies can easily access publicly available datasets through established channels—this approach poses no significant challenges beyond time and cost considerations. As we move toward the era of true Artificial General Intelligence (AGI), a common vision is for models to achieve self-iteration and self-learning capabilities, enabling them to train themselves autonomously.

Regarding this aspect, based on the current data, do you think simulation data is viable, or does real-world data maintain the highest quality? If real data remains essential, could this limit the AI's capabilities to the level of human historical experience, given its reliance on authentic historical data?

Or can we still overcome this limitation by utilizing simulated data, experimental data, generated data, and other methods to enable the model's capabilities to surpass all previous human achievements?..... I believe this can be surpassed. There are two key areas where it can exceed human performance—for instance, in Go: AlphaGo played a move never seen by humans, demonstrating that it has certainly surpassed humans within certain limits. However, it may also have inherent limitations.

However, this limitation is not immediately apparent to us. We believe that, in general terms, it can be overcome by leveraging the existing knowledge we already possess and can articulate. The question then arises: should this approach rely on real data or simulated data? Will it work? There are indeed multiple possible methods. Thank you for your time; I would also like to follow up on the earlier question regarding AI infrastructure. What is your second question?

Alright, let me briefly and quickly repeat it. My question is this: regarding AI infrastructure, I believe that everyone is currently investing hundreds of billions of dollars in computing power. In this regard, we are confident that China will achieve its goals in hardware development in the future—there may come a day when highly efficient computing power becomes available. However, this could still be a bottleneck during implementation. So, will there be two parallel paths moving forward?

On one hand, as models undergo continuous capability upgrades, their computational approach evolves from brute-force processing to intelligent computation, leading to a gradual reduction in the computational demands and resource consumption per model or task. On the other hand, advancements in hardware performance—such as those in GPUs—accelerate iteration speeds and enhance single-card efficiency. What phenomenon emerges during this process?

Could it be that after building a 10,000-card cluster and purchasing H200 or 960 GPUs, within two to three years they become relatively less powerful computing resources and eventually turn into outdated components? NVIDIA GPUs can generally be depreciated over five years, while Huawei GPUs are depreciated over at most three years. The Huawei 950 is still performing quite well this year; I think it'll hold up reasonably well next year, but beyond that, it might really consume excessive power.

The lifecycle of Huawei cards is definitely shorter, as they were already two years behind NVIDIA from the outset. However, I don't think the gap is that significant. If the B200 models are available for purchase now, I believe they're all cost-effective. For Tencent, if Alibaba can acquire them, we'll need to consider the volume; if they can be purchased at a reasonable price, they're undoubtedly advantageous. But this isn't about calculating costs—I'm confident they won't be available. We understand this: our computing power is lagging behind, and this is a fact. This reality can be addressed through three approaches.

The first aspect involves accepting model limitations: we can only use models that are smaller than theirs, and the reduction in size directly impacts training efficiency. We must tolerate this degree of model underperformance and smaller model sizes. However, there is a benefit to this limitation—it provides more time and allows us to leverage existing technical expertise, enabling the application of sophisticated approaches.

【02:53:59】

Thank you. Therefore, the gap between us and the United States may range from 12 to 18 months, or more precisely, from 6 to 12 months. Simply put, it means lagging behind the United States by two years while achieving the same outcome using only one-twentieth of its computational resources. The core premise is that despite a two-year delay, the task can be accomplished with just one-twentieth of the required computing power.

In the future, we need to rewrite this narrative: using only a fraction of their computing power while significantly shortening the timeline to just six months or three months—I believe this is our goal. Moreover, we could even surpass them in certain aspects. However, given the substantial disparity in overall computing capabilities, achieving comprehensive superiority is unrealistic; yet in key areas where trade-offs are necessary, surpassing them may be feasible. Thank you, Mr.Liang.

Full of confidence, let's work together. Time is also available for other participants—thank you. Thank you for your recent sharing. I have two quick technical questions. During our discussion on the technical approach, we mentioned that the core challenge at this stage is continuous learning, which is also a current focus in international research (known as recursive improvement). Could you please explain what the biggest technical difficulty currently is, and when do you think it can be resolved?

This is the first question. Regarding the second question, as you just mentioned, we should first address continuous learning before moving on to artificial intelligence. I'd also like to understand the underlying technical rationale behind this approach—does it imply that once continuous learning is resolved, DeepSeek will subsequently develop general intelligence? These are two points I'd like to clarify. The technical challenge lies in the fact that we haven't yet identified a truly effective methodology.

The world has yet to find an effective solution; everyone is still exploring. Therefore, we remain in this exploratory phase, uncertain who will ultimately discover the next approach to solving this problem. We have numerous ideas and concepts that currently appear promising, but none have been fully implemented. First, we place great emphasis on a key principle: when training our next model version, we aim for it to directly support our own development efforts.

It enhances DeepSeek's efficiency. Our model primarily aims to improve DeepSeek's operational performance, enabling it to provide greater support when developing subsequent versions. Simply put, our primary goal is not for others to use the model effectively, but for us to use it efficiently—first and foremost, for our own benefit. Once we find it useful, developing the next version becomes significantly faster.

Many of us internally hold this view: First, it must be useful for ourselves—primarily for our own needs. Moreover, this is the fastest way to achieve AGI. If it works well for us, it likely will work well for others; but we must first ensure its effectiveness for ourselves. This perspective may sound unconventional, yet it reflects the mindset of many. My point isn't that users should prioritize usability; rather, I believe focusing on what benefits us most will significantly accelerate our progress toward AGI.

The underlying logic of this narrative is to facilitate our achievement of AGI – but first, it must help us achieve it. We genuinely require this assistance to realize AGI. It is now absolutely certain that artificial intelligence is essential for achieving AGI. Although it hasn't yet operated autonomously and still works in conjunction with humans, its utility is already significant. The second challenge, as mentioned earlier, is addressing the issue of continuous learning before progressing toward general intelligence. Is this your anticipated roadmap for the future?

I'd like to understand the underlying technical rationale behind this approach: why must we first address continuous learning before progressing to general intelligence? What are your subsequent insights on this matter? Overcoming the challenge of continuous learning can significantly accelerate our R&D progress. Once we resolve this issue, achieving general intelligence becomes much more straightforward. With AI's support, its capabilities would be remarkably robust if it could learn continuously.

The current limitations of agents stem from their inability to engage in sustained learning effectively. If continuous learning could be fully implemented, artificial intelligence would become immensely powerful, significantly enhancing the efficiency of our research endeavors. Once continuous learning is achieved, general intelligence would become readily attainable and easily applicable. This is precisely the outcome we aspire to see – one that reduces effort and simplifies tasks considerably.

Otherwise, developing general artificial intelligence manually would be extremely labor-intensive and resource-consuming—a data-and manpower-intensive process with poor cost-effectiveness. Thank you for sharing. For further details on this topic, please refer to the chat group. How long do you estimate it will take to develop AGI? Can domestic hardware catch up within that timeframe? This discussion took place in the Zoom meeting chat window.

Alright, I've reviewed this. Regarding the Huawei 950, Huawei currently provides us with 16,000 SIM cards, which should be publicly disclosed.

It should be one order of magnitude less than that of major internet companies. Huawei can only provide us with this amount, as the price is not cheap either. Major internet companies have even greater aspirations and require more from such products. For us, we can purchase some non-compliant cards. Therefore, the purpose of buying the Huawei 950 is still to help Huawei build a robust ecosystem. A Huawei 950 with 16,000 cards is equivalent to only a B-series card with 4,000 cards.

Therefore, it is not a very large volume and has limited significance. It is insufficient for training a next-generation model; it only meets the requirements for training our current generation of models and falls short for training future models. However, it enables Huawei to complete this task first—specifically regarding the Huawei 950. Furthermore, how much longer will AGI take to develop? Can domestic hardware catch up within this timeframe?

I believe that in the field of AI, China should be able to match international standards within one or two years, or even develop models that rival foreign ones this year. Given the current approaches and paradigms, achieving this is not particularly difficult—it should be feasible this year. However, this level of progress does not yet constitute AGI.

【03:05:48】

At the very least, I believe it must be capable of continuous learning. Can domestic hardware catch up by now? I think it may take several years for domestic manufacturers to achieve this. First, China needs to address the ecosystem issue, as it fundamentally relates to confidence building. Once the ecosystem is established, the production capacity problem can then be tackled – I believe these challenges can be resolved gradually. I don't expect we'll still be stuck with production capacity issues in the next five years.

The current bottleneck undoubtedly lies in production capacity; I believe this issue will persist throughout this year, next year, and the following year as well. However, five years from now, that may not necessarily be the case—I remain quite optimistic. The second aspect concerns future organizational structure and personnel planning scale. Initially, our organizational structure was highly fragmented due to a lack of clear framework. As our workforce continues to grow, significant adjustments are imperative.

I can only say that significant changes will be required here, but it's difficult to articulate them all at once. Ultimately, we'll need to categorize departments accordingly: some will require a more rigorous hierarchical structure, while others may maintain a flatter, more flexible setup. As our workforce grows, we'll make these adjustments. In fact, we need to implement them immediately, as I'm already working on this change.

Without making this adjustment, many tasks cannot proceed effectively. Indeed, numerous departments require well-defined organizational structures. Another common question is: which version preceded CV? Currently, the GCV4 version released online remains relatively preliminary, and significant improvements in functionality still require time. For me, a suitable release schedule would be approximately every two to three months.

The next release is likely scheduled for late June, around the end of April. barring any unforeseen issues, each subsequent version should improve progressively. At the 50B activation scale, I believe the final outcome will not differ significantly from current open-source implementations—particularly in inference speed and performance. However, compared to its large-scale model (the only unreleased variant), there remains a notable gap.

The gap lies in the fact that our current activation parameters simply cannot achieve this level of performance, which would require a model with significantly larger parameters—perhaps around 150 billion parameters. Given our current training progress, I estimate we could begin training by the end of this year at best, or at least by next April; the difference remains substantial. This is precisely the gap compared to OCE's capabilities.

Hello, I have a question: you often mention that AGI is developed through a gradual process rather than an abrupt leap. So could it be understood as a process without a tipping point? Indeed, there's no clear threshold, but the development is nonlinear. Currently, we largely subscribe to the narrative that AI can accelerate AI research itself.

In other words, the relationship is not linear; since you can use AI to accelerate your research, it may become nonlinear later on. So, based on my understanding of the previous conclusion, scaling language models sufficiently to this extent is entirely adequate. I don't see any upper limit to how far language models can scale—neither does our current level of intelligence nor that of the United States.

I understand. My curiosity stems from your observation that for the United States, implementing this 800-billion-parameter model requires significant computational resources—it's capable of training but impractical for widespread use due to its high cost. What particularly intrigues me is that human language ability has only emerged over the past 100,000 years, having evolved over approximately 3.7 billion years prior.

However, when it comes to training AI, while the order can potentially be reversed, the process ultimately still leads to what might not even be called a "world model" —it must inevitably reach the realm of physical models or embodiment, right? That's beyond this fundamental threshold. Indeed, I believe embodiment remains essential; it's the ultimate goal. For our company, the natural endpoint is undoubtedly embodiment. After all, for an average person, their primary need isn't a computer, is that correct?

For ordinary individuals, their daily needs for food, clothing, housing, and transportation do not require a computer. What they truly need is embodied intelligence to address specific human labor demands. If the goal is to meet these labor requirements, then embodied intelligence is indispensable. I understand this point.

Phasing-wise speaking, reaching a certain level doesn't necessarily mean hitting a critical point—it could involve self-evolution, or at least significant self-improvement, or approach the stage where SV performance rises to that specific point. I'm curious whether the initial deployment of this state of AI might differ from current practices. We envision that even without physical embodiment, our definition of AGI—or what we hope AGI can achieve—would still be distinct.

It can assist me in refining the next version of the model. Similarly, once the embodied system is developed, we expect it to evolve further by creating the next iteration of the robot. I'm particularly curious—based on my reading of previous interviews with DeepSeek and related discussions—that taste and intuition play a crucial role in making pivotal decisions regarding research directions and strategic choices, rather than relying solely on engineering optimizations.

So if AI can evolve autonomously in the future, will elements like taste, intuition, and similar qualities remain important—or what exactly would become crucial? Currently, AI doesn't lack taste or intuition; what it lacks is the ability for continuous learning. There's no issue with AI's sense of taste or intuition. When asked to write an article, I believe its judgment and intuition are fully adequate. There are a few other points mentioned earlier—I'll take a look at them. I see several issues listed on the screen, but they're not visible here.

Mr.Liang, I've left a question on the screen – let me restate it for your reference. My inquiry concerns "continuous learning," which you and many researchers have noted remains an unresolved research challenge. Additionally, regarding that coding agent, particularly its integration with MILES...

【03:17:47】

Reaching scaling targets at both the Office and MIS levels is a well-defined objective. Given an unresolved research goal alongside a clearly defined scaling objective, how should research resources—particularly human resources among researchers—be allocated to achieve optimal balance and effectiveness?

Model Office is a relatively well-defined objective, whereas Model MIS remains somewhat ambiguous—I would describe it merely as a goal. As for Model Office, I believe it's quite clearly defined. CoT does not consume resources.For any collaborative research project, he doesn't consume cards—he only needs a minimal number; what he truly requires are ideas. Nor does he require human resources, as no专人 needs to be constantly involved.

This isn't a single project per se; rather, it requires many people to engage with this concept. Therefore, no specific resources need to be allocated—it simply doesn't require any. What you actually need are resources for training models, developing models, and conducting efficiency experiments on model development. As mentioned earlier, the resource consumption for both human effort and computational resources is minimal. That's why we call this process "lottery." The barrier to entry is very low—anyone can participate—but what results one achieves may depend more on innate talent than other factors.

Therefore, there is no need for us to allocate specific resources here. What sets us apart from other companies is that we take the time to discuss and reflect on this issue, treating it as a matter of great importance. Within our organization, it is indeed a critical concern that requires dedicated consideration, though not demanding significant resources. Additionally, I would like to highlight another issue: the hallucination problem in large models significantly impacts user experience.

The issue of hallucinations can also be addressed, though it remains a long-term challenge. It could potentially be resolved through improved post-training approaches—a solvable and improvable problem that simply hasn't received sufficient attention. To me, hallucinations are indeed an issue, but we tend to attribute them primarily to product limitations. We do address it, but not as a top priority; prior to that, there's also the issue of labeled data.

In terms of data annotation, this is related to our capital investment. Given the structure of our capital investment, we cannot afford the high costs associated with producing such a large volume of high-quality annotated data. The cost of data annotation in the United States is no different from that in China. There is no cost advantage for China in annotating data, especially for high-end data, which makes it difficult for us to invest in data annotation at the same level as in the United States.

This task is extremely challenging in China due to the exorbitant cost of labeled data—whether it comes from external sources or our own efforts, the burden is significant. Therefore, we currently adopt a dual-track approach: while we don't completely abandon data labeling, we prioritize tasks with lower costs over those with higher costs. In fact, half of our company's workforce is dedicated to data labeling; this includes half of our core researchers and key personnel. We have thus focused our efforts entirely on this critical task.

To address the issue of AI at this stage, it relies entirely on labeled data. Just take a look – it's all data. Mr.Liang, thank you. Your remarks were particularly insightful; we truly appreciate them. Thank you. I would like to ask a few questions. First, as you mentioned earlier, China's models are undoubtedly more efficient than those from the United States. You also highlighted other aspects where they may potentially surpass American models in the future. In what areas do you think we could outperform the United States, whether in intelligence or elsewhere?

I believe that in many aspects of user experience, we might actually perform better than the United States. Not to mention our own user experience—which I find quite satisfactory—I think we may not necessarily lag behind the U.S. in terms of overall user experience. Regarding product quality and capabilities, we may also be on par with the U.S., and costs are likely lower as well. Therefore, China could still maintain competitiveness. As for other structural advantages, I don't think they exist either.

However, in terms of both cost and product development, I believe there are indeed certain structural advantages. The cost advantage is straightforward: since they don't need to develop these capabilities themselves, they naturally neglect them—unlike us who prioritize this aspect. While we consider it crucial, it's not important for them. The same applies to product development; through collaboration with multiple companies, their product capabilities remain robust. Therefore, I think these two areas represent significant structural strengths.

Regarding your second question, you mentioned that post-training requires relatively high investment costs—both Anthropic and OpenAI have allocated substantial funds to this area. Following this round of financing, do you expect increased investment in post-training? The main gaps lie in high-quality data annotation and AI research itself. We will undoubtedly increase our investment, but high-quality data annotation typically does not involve significant capital outlay.

The main bottleneck in high-quality data annotation, I believe, is time—it simply requires more time. For organizations like OpenAI, international players, and Anthropic, they started earlier and have greater capital resources and access to resources. In contrast, China has only begun this effort in the past six months, so we'll need significantly more time.

This has little to do with capital investment, as even without additional funding, the existing capital is sufficient for rapid expansion. However, this growth rate has its limits—the bottleneck lies not in immediate recruitment of more personnel, nor in financial constraints or operational bottlenecks. Nevertheless, it remains a process of rapid expansion. Therefore, we believe that high-quality data management will be well achieved within a year, which is likely to be achievable domestically.

I don't think the situation is particularly dire, but it certainly takes time. Thank you. Regarding the third point, we've observed that Anthropic uses its own models to develop products, having launched numerous solutions in vertical sectors such as finance and law.....

【03:29:17】

Even if we plan to move into the healthcare sector in the future, do you think we will consider such vertical applications at some stage? I haven't thought it through thoroughly yet; we still don't have a clear understanding of what our domestic business model will look like or what the most viable path would be. The domestic situation may differ from that abroad, and I believe it's still quite difficult to predict how things will develop domestically at that time.

Given the current situation in China, I believe the most sensible approach is to focus entirely on developing a general-purpose Agent, while giving lower priority to other Agents—such as those for finance or healthcare. Coding should be our top priority, as it enables the development of versatile agents alongside specialized vertical agents. At this stage, the Coding Agent remains our foremost focus.

That's explained very clearly; thank you. We also have another question we'd like to ask. In fact, when developing DeepSeek, we greatly admire your approach—you've consistently pursued it with a purely scientific methodology.But now this industry has indeed entered the realm of capital markets and embarked on a path of monetization. You are also an exceptionally responsible individual, demonstrating great accountability both toward your colleagues and investors.

So, regarding the balance between pure research focused on AGI development and engagement with capital markets, you will undoubtedly continue to operate in the capital market sector while also maintaining a public equity presence. How do you envision achieving this balance going forward? What are your considerations? I believe it's entirely achievable – both approaches are essential. For instance, if we generate several hundred million US dollars in B2B revenue this year, combined with a user base in the B2C segment, that alone would establish a solid commercial foundation.

By next year, we will generate revenue from our B2B segment. If demand continues to grow, the company will be close to achieving net profit – possibly even reaching it. This would mark the end of a purely capital-intensive phase, so I believe there are significant opportunities for further strategic moves. Even in the worst-case scenario of selling APIs, the business could still sustain operations as a publicly listed company.

If we assume that without further technological advancements, our technology would stagnate, then ultimately we could focus solely on selling APIs and optimizing these services – I believe that would be sufficient. Therefore, I remain confident; it's not actually that difficult. After all, we're operating in a high-leverage environment within a rapidly evolving field, which makes the challenge less daunting. Our goal is to pursue greater ambitions while maintaining achievable baseline performance. Thank you for your excellent presentation.

That concludes my questions; thank you for sharing. Regarding some points mentioned earlier, I would like to ask a few additional questions. I believe the most significant distinction between DeepSeek and other companies lies in our unique organizational structure. This structure is not only aligned with our objectives but also requires careful consideration of its boundaries and efficiency. From a macro perspective, I wonder if there exists a suitable benchmark or model for our organizational framework to emulate.

Historically, what form might Bell Labs have taken that would be considered ideal? Or perhaps we ourselves don't have a clear ideal; it largely depends on our ongoing exploration. In this new era, we must rely solely on our own efforts to explore. After all, our organizational structure is certainly different from those of several American companies, isn't it? Even the three companies themselves are distinct, yet they at least approach their endeavors from a commercial perspective.

I would like to reiterate my question regarding organizational structure. Firstly, we have no model to emulate. Every decision we make is grounded in reality—we act pragmatically, base our choices on actual circumstances, and determine the appropriate course of action. Therefore, this approach is a product of its time or a reflection of prevailing conditions; it does not result from mere imitation. In other words, under these specific circumstances, what I consider the optimal solution may indeed be this one, or rather, this is the path we have chosen ourselves.

Every step we take has been carefully considered and deliberate; the choices we made were indeed based on thorough analysis. Our decisions weren't driven by observing others' approaches, but rather by weighing pros and cons. The same principle applies in the future—we won't simply imitate anyone's methods. I believe we differ from Bell Labs in that our approach explicitly doesn't require commercialization, whereas we are committed to achieving commercial success.

Ultimately, our priority is survival—we are, after all, a company, and the government won't provide any funding. While we can pursue ambitious goals, we must fundamentally focus on how to sustain ourselves as a business entity. The B2B segment is therefore crucial, as it may become essential for our long-term viability. However, it's not a top priority at present, as it currently represents a cost burden. I believe this situation differs significantly from that of Bell Labs.

At its core, we remain a company. Throughout history, many companies have pursued objectives beyond profit, yet this does not render them non-corporate entities. Numerous companies have achieved remarkable success precisely because they embraced such non-profit-driven aspirations. In fact, these pursuits not only did not hinder their commercial success but rather enhanced it significantly.

At its core, we remain a company; we simply need to determine what revenue streams to pursue, when to generate them, how much to earn, and on what basis—essentially making strategic trade-offs. Mr.Liang, I have two quick questions for you. First, you mentioned MILOS as a potential goal, which may not be fully defined yet but is definitely the direction we're moving toward. Second, you also referenced activation parameters, such as.....

【03:41:02】

The next generation is likely to fall within the range of approximately 150 to 250 B. In this scenario, do you believe 150–250 B refers to performance benchmarking against O4.7 or possibly other benchmarks? That's the first question. The second question, as you mentioned earlier, concerns our current use of compilation languages other than CUDA throughout the inference pipeline.

I understand that, given our reliance on the NVIDIA ecosystem, we previously relied heavily on solutions like PTX. Now that we're adopting technologies such as TileLang you just mentioned, could this significantly reduce our inference efficiency—or at least temporarily? I'm curious how you view the efficiency trade-off associated with these compiler-driven changes; or do you believe they ultimately represent a complementary approach that enhances overall performance in the long run?

This significantly boosts efficiency—indeed, a substantial improvement. So, does it have any negative effects? No, it doesn't; it dramatically enhances efficiency. This presents a genuine opportunity. Whereas previously you were dependent on the CUDA ecosystem, now we can abandon it and adopt a simpler approach: TileLang. As an advanced language, it allows rapid programming with minimal code required, enabling complete code rewriting if needed. Understanding?

So these two aspects are essentially the same. As you mentioned, the significant opportunities brought by AI are not shortcomings that need to be addressed in the short term; rather, they represent major opportunities for technological advancement. This isn't solely about AI itself—after all, we have another project where we're using AI to develop TileLang. Is it faster? Currently, all TileLang code is written manually, but it's already much quicker than writing it in CUDA. Got it?

So even regarding the underlying hardware performance, do you think it has no impact?A loss of 1% to 2% is acceptable in my opinion. Okay, I understand. Thank you for making it clear. Do anyone else have questions? No issues; let's proceed with this for today. Thank you all for your time. Thank you, Mr.Liang. Goodbye.

Editor: Pang Tong · WeChat: pongtom

This article was compiled by **Pangtong General AI.**