Midfield contest between large models: who can get companies to use large models first

*Note: These are Jeffrey Ding's informal and unofficial translations -- all credit for the original goes to the authors and the original text linked below. These are informal translations and all credit for the original work goes to the authors. Others are welcome to share **excerpts** from these translations as long as my original translation is cited. Commenters should be aware that the Google Doc is also publicly shareable by link. These translations are part of the ChinAI newsletter - weekly-updated library of translations from Chinese thinkers on AI-related issues: https://chinai.substack.com/*

Source: Leiphone
Author: Zhang Jin

Date: September 6, 2023
Original Mandarin: https://mp.weixin.qq.com/s/1VhOEGnXPfvM6D0Uy8ea5A

"Large model makers need to not only give people fish but also teach them how to fish."

With the accumulation of capital and discourse, the craze for big models has swept the country for more than half a year, and suddenly we have entered the era of large models available to everyone.

But in fact, before August 31, each large AI model product was still in the trial stage and had not yet been officially qualified to provide services to the public. The early warm-up was at its peak until August 31 when big news hit. The first AI large model products of eight Chinese enterprises/institutions, including Baidu's Ernie Bot and Chinese Academy of Sciences's "Zidong Taichu large model," had passed the filing with the "Interim Measures for the Management of Generative AI Services", which means that they have officially obtained qualifications to provide services to the public.

According to data from Qimai, on the first day it was opened for download on August 31, the "Wenxin Yiyan (Ernie Bot)" app was downloaded an estimated 313,000 times in the Apple App Store, while the Zhipu Qingyan App was downloaded 3,832 times. According to official data released by Baidu, within 24 hours, Ernie Bot responded to more than 33.42 million questions from netizens.

The comprehensive "opening" of AI large models is a critical moment for the development of domestic large models. The above eight large model products have first-mover advantages.

The direct confrontation between Chinese large-scale models has officially begun. Previously, domestic large-scale model products similar to ChatGPT all focused on "conversational interaction.” Competition among various companies over the strength of their large models was simply focused on quantifiable dimensions such as accuracy or benchmark rankings in question answering. Most of the conversations between consumer-side (C-side) users and AI trended toward casual chatting, rather than creation. On the other hand, C-end users’ willingness to pay is low, and to-C’s general large-model products cannot replicate ChatGPT’s successful model in the short term. Under the siege of homogeneous products, they are forced to join the money-burning game of resource competition. Looking back, To B customers have high willingness to pay and great demand. As the market returns to rationality, the choice of To B for large models has almost become a tacit paradigm in the industry.

Therefore, in the future, the real arena for large model competition will mainly focus on serving B-end users. At present, large companies and start-ups are refocusing their attention on the industry.

Under the glitz, we began to dig out which large model producers are thinking about the real value of large models - in addition to the excitement of AI intelligence brought by ChatGPT, what can large models bring to the industry? And in the era of large AI models, what does the industry need? How can their large model capabilities be integrated with industry? We are also wary of whether this is another “showy but not practical” AI technology competition between major manufacturers.

Only when the tide recedes will we know who has been swimming naked.

1. The gap between technology and implementation in the era of large models

Some media reported that as of July this year, China had released a total of 130 large models, surpassing the United States and entering the first echelon of large models. Large models are deemed a new generation of infrastructure, but on the industrial side, many companies do not really use large models. The main reason why is because (government) policies have not opened up, but the enthusiasm on the industry side is very high, even more so than on the C side.

Wang Feng, CEO of a company that specializes in "consumer content + marketing services", told Leiphone that there is now a saying in the industry that 90% of content on the Internet in the future will be generated by AI. This means that the challenges and impact of large models on their business are very large, so they must respond early.

In order to cope with this impact, they transferred some people from the original teams at the beginning of this year to establish an AI Lab team to explore and try out large-scale models. In Wang Feng's view, the learning and adaptability of general large models is still very strong. Everyone has seen how amazing it is, but it also has certain limitations. For example, when we talk to GPT or Chinese models, we ask them how to choose tea leaves. It will only generally tell you what to pay attention to. Many times, the output content does not directly help consumers make decisions.

Therefore, in Wang Feng's view, large models are not omnipotent. One of their limitations is that they cannot solve vertical problems. This is one of the reasons why many industrial companies currently fail to use large model capabilities.

However, pre-training a basic large model from scratch requires a very large capital investment, which is unrealistic for companies like them, and it is difficult for many companies to do it even if they have funds, data, computing power, and know-how, and maintenance capacity, etc. These are all barriers, so industry can only seek cooperation with existing large model producers on the market.

Sellers of domestic model capabilities can be divided into two categories: one consists of mainly large companies such as BAT as well as start-ups focused on large models; in addition, there are large model intermediaries, mainly founding teams that develop application services based on large models, including providers of underlying computing power and frameworks, and even third-party companies that provide fine-tuning of large models.

After some research, Wang Feng decided to train (fine-tune) his own model: building on top of large models, he poured his professional knowledge of the industry into further training and fine-tuning. After training, he deployed it locally for private deployment, and then solved the corresponding vertical problems.

But in the process, they found that they didn't know which model to use: There are currently too many large models published in China. If they want to verify all the models and compare the effects of the models one by one, then the labor cost will be very high.

In fact, B-side enterprise customers often bring their own scenarios and data, which is the best testing ground for the implementation of large models. But when the theory lands on reality, Wang Feng's dilemma is also a common problem faced by many B-side companies.

Therefore, based on the above reasons, there is currently a high wall between domestic large models and the industry. The capabilities of large models on one side of the wall cannot be released, and the digital needs on the other side cannot be met. Moreover, large models trained based on open data sets are not good at professional knowledge, and enterprise users who have a grasp on industry data cannot participate in the construction of large models.

02 To-B solution of platform mode

In order to tear down this wall to the greatest extent, to release large model capabilities to the industry side, and allow enterprise users who have a grasp on industry data to truly participate in the construction of large models, Baidu AI Cloud has proposed its own solution: “Qianfan” Large Model Platform + Solution +AI native application.

Based on this, in order to help enterprises and developers quickly retrain based on basic large models and build enterprise-specific large models, Baidu AI Cloud launched the Baidu AI Cloud Qianfan large model platform. On the Qianfan platform, users can directly call 42 large model services, including Ernie Bot. You can also develop, deploy, and call your own large industry models on Qianfan.

The Qianfan large model platform provides enterprises with a full-process tool chain and a complete set of environments for large model development. Users can complete all aspects of large model development, training, deployment, and application development. The upgraded Qianfan 2.0 tool chain covers the entire life cycle of large model development. This includes: data management, model training, evaluation & optimization, prediction services and prompt engineering. It helps enterprises to efficiently develop and deploy large model applications end-to-end, and continuously lower the threshold of large model technology.

When Baidu AI Cloud conducted market research in the early stage, it discovered that many corporate customers are becoming more and more professional and rational in their selection strategy for foundational large models.

It used to be that when corporate customers were trying to comprehend the basic capabilities of large models for a large model producer, they just looked at the rankings. Now when they choose a large model, they have to optimize it based on their own scenarios and data, and they will consider many things, such as the performance of the model, the efficiency of development, and usage costs. The so-called usage costs, such as resource utilization, how big the model is and how many resources it requires; how the performance is; fine-tuning according to enterprise tasks and the cost of tuning. These are all factors that enterprises consider when choosing a basic model.

Today, everyone in the industry has begun to seriously consider how large models can bring value to themselves as infrastructure, instead of just joining in the fun like in the early days.

Based on the aforementioned B-end users’ understanding of large models, in order to meet the demands of different users, the Qianfan platform has also integrated 42 mainstream large models (from both Chinese and international labs), making it easier for users to choose according to their own business segmentation scenarios.

The above-mentioned companies that can clearly understand their own large model application scenarios and develop large models to varying degrees are mostly concentrated in the Internet

industry, and their knowledge of technology and industry is usually at the forefront. Therefore, as long as they are provided with a good and complete tool chain and a complete set of environments, they can "self-service" and meet their own large model needs on the Qianfan platform. However, in some traditional industries, their scenarios are complex and their understanding of large model technology is insufficient. Large model producers must go deep into the industry and accompany enterprises to sort out large model application scenarios and make use of large model capabilities.

Therefore, Baidu AI Cloud has reconstructed solutions based on large model capabilities in the four major industries of digital government, finance, industry, and transportation.

1. Baidu AI Cloud Digital Government Solution "Jiuzhou" [九州] has comprehensively enhanced the large government model, including knowledge enhancement, retrieval enhancement, cognitive iteration and security policy enhancement.

2. Baidu AI Cloud Financial Solution "Kaiyuan" [开元], a retail analysis assistant that can assist financial managers to grasp customer dynamics in real time, provide accurate business insight analysis, intelligently recommend key tasks based on business goals, and generate marketing rhetoric for thousands of people through large models.

3. Baidu's AI Cloud Industrial Solution "Kaiwu" [开物] is upgraded based on the Wenxin large model. This new Kaiwu will realize the leap and improvement from "production line intelligence" to "enterprise intelligence" to "industry chain intelligence".

4. Intelligent transportation solution ACE 3.0. After being redesigned and optimized based on the large model, the traffic organization plan achieved better "diagnosis and treatment effects" and covered a wider area. For example, in the past, it was difficult for the traffic police department to accurately identify debris spills (trash, plastic bags, bottles, etc.), but large models have greatly improved the accuracy of identifying such incidents.

The above four major solutions are more of Baidu AI Cloud's "proofing" for customers and ecological partners. Because the Qianfan large model platform expects a prosperous ecosystem in the future. As more and more users are gathering on the platform, it is inseparable from customers and the co-construction of ecological partners. The platform looks forward to developing more rich solutions and products.

In addition to vertical solutions for the four major industries, Baidu AI Cloud also saw that a large number of customers have common application scenarios and business needs. They sorted these into common needs and officially launched them for cross-industry applications. Family is an AI native application for common industrial application scenarios, including digital human Baidu AI Cloud XiLing, enterprise search engine Zhenzhi, safety production intelligent assistant Du'anan, etc. These applications cover three key scenarios: "marketing services, office efficiency improvement, and production optimization." Baidu released 11 AI native application

products at once to meet the common needs of different industries and accelerate the large-scale implementation of large models.

They are also all "model rooms", which means that in the future, enterprise users will also have the opportunity to develop such AI native applications on the Qianfan platform. But at the same time, it is not just a model room, because it is also a product that meets the needs of common scenarios in the industry.

Baidu AI Cloud has released its own Qianfan large model platform, industry solutions, and AI native application Family. From tool platforms to products and partners, it promotes large model applications in an all-round way and accelerates the implementation of large model applications.

3. Baidu, on what basis?

In this large-scale model competition, Baidu AI Cloud is not the only one adopting the platform model to build To-B services.

In March this year, Baidu AI Cloud Qianfan large model platform was launched. According to official news from Baidu Smart Cloud, the number of active companies in Qianfan today has exceeded 10,000, covering more than 400 scenarios in industries such as manufacturing, energy, government affairs, and transportation. It truly allows many companies to experience the "intelligent power" of large models.

The most direct efficiency improvement is that in the past, if a customer wanted to see the effect of a large model, just debugging, verification and evaluation would require the algorithm team to invest a week. Now on Qianfan's one-stop tool chain platform, customers can run through a model and see the effect on the same day, and quickly enter the large model application development stage.

According to Leiphone, Baidu AI Cloud has done several things right:

1) The Qianfan platform has the largest number of large-scale models in China, and model inference costs can be reduced by 50%. The platform has access to 42 mainstream large models such as the full series of Llama 2, ChatGLM2-6B, RWKV-4-World, MPT-7B-Instruct, and Falcon-7B. Users can choose one or more large models to apply according to different business scenarios, thereby achieving “model freedom.” This is the effect that many manufacturers who want to build large model platforms want. According to Leiphone, there are only a few third-party large models on some large model platforms, and the rest are all their own large models.

These 42 large models have been strictly selected by Baidu. They have not only passed the three-dimensional assessment of model performance, model security, and commercial availability, but also made model security enhancements for all connected third-party models,

which not only ensures the content security of Wenxin's (Ernie's) large models, but also ensures the safe output of third-party large models.

But Qianfan doesn't just meaninglessly gather large models together to serve as a "convenience store". Instead, Qianfan makes secondary performance enhancements for each connected large model, including: a. Chinese-language enhancement. The Chinese language has been enhanced for the mainstream foreign large models. For large foreign models like Llama2, it used to be possible to communicate in English, but now it works just as well in Chinese. b. Performance enhancement can comprehensively improve training and inference performance. The overall throughput of training LlaMA 2 can be increased by 25%, and the inference performance can even be increased by 2 times. c. Provide context enhancement for open source models to meet the reasoning needs of various long context scenarios including knowledge enhancement, long-term memory enhancement, and document-based question answering.

2) Qianfan platform has initialized 41 data sets and 10 selected application paradigms. When Baidu AI Cloud was serving customers, it found that a large number of customers found it too troublesome to access and manage data when doing model fine-tuning. They hoped that some good data sets could be preset on the Qianfan platform in advance. In view of this demand, the upgraded Qianfan Platform 2.0 has initialized 41 high-quality, industry-specific data sets. Users can complete fine-tuning and improve model effects by clicking a few buttons.

After selecting the model, the next step is application development. In order to further improve the efficiency of application development of large models, the Qianfan platform provides 10 featured application paradigms such as knowledge Q&A and customer service dialogue.

IDC predicts that China's AI large model market will reach US$21.1 billion in 2026, and AI will enter a critical period for large-scale implementation.

Every company that has the ability to independently research large models wants to be an infrastructure operator in the era of large models. After a few months, whether large models are really the smart operating system of the AI era can only be determined by industry movements. When it comes to societal infrastructure such as water and power grids, no matter how many existing industry players there are, only a few will succeed in the end.

At present, domestic large model makers mainly include companies such as Baidu, Alibaba, Huawei, Tencent, and SenseTime, as well as research institutions such as Beijing Academy of Artificial Intelligence and the Institute of Automation of the Chinese Academy of Sciences. At the same time, chip manufacturers such as Nvidia have also entered the market.

According to Leiphone, Baidu has obvious advantages in the competition among infrastructure operators in the era of large models.

The first is the opportunity as an AI cloud vendor.

For many companies making native AI applications, a big pain point is the high cost of inference. Because inference is very resource-intensive, you need to buy a lot of machines, and it is difficult to control the quantity. If you buy too much, it is a waste, and if you buy too little, the inference capabilities cannot keep up with user volume. This is a big opportunity for cloud vendors, because the characteristic of the cloud is elastic resources. Cloud vendors can provide elastic inference resources for these companies that make AI native applications through scale and technology optimization.

*****Excerpted three passages that fill out more details about cloud-AI integration**

At the same time, another advantage of Baidu is that Baidu has accumulated AI technology and explored AI scenarios earlier than other manufacturers.

As early as the beginning of 2010, Baidu began to explore artificial intelligence, and since then has continued to increase investment in research and development in the AI field. At that time, few people could foresee the future development of AI. However, Baidu's R&D investment of hundreds of billions shows Baidu's obsession with technology. It can also be said that Baidu made the right bet in the direction of AI.

Next, the bottom layer of Baidu's own business is driven by AI. Baidu can usually implement the latest AI technology quickly and on a large scale internally before other manufacturers. When facing users, Baidu's large-scale model capabilities are naturally convincing. Today, products such as search, Baidu Netdisk (Wangpan), digital human, and BaiDuHi (company knowledge management services) have all been reconstructed through generative AI.

Another advantage of Baidu is its deep penetration into the industry and its involvement. Since the release of the Qianfan platform in March, the Baidu AI Cloud team has been exposed to more than 400 application scenarios. A lot of manpower has been invested in supporting each scenario to understand the customer's scenario, what method to use, what the performance is, and how to improve it. After receiving market feedback, the team quickly iterates the tool chain on the Qianfan platform based on this feedback, forming a positive cycle to continuously optimize the platform.

Cundao Yi, chairman of Beijing Baolande Software, said that how to build large models well in vertical industries is currently a problem that all companies must consider. Most companies are not as capable as Baidu in building large models. How to effectively combine Baidu's large model capabilities with corporate business to create stronger value is a very good opportunity for companies to overtake at the corner.

Today, we can feel that large model technology changes too fast. For enterprises, the cost of keeping up with the trend is very high. Therefore, Qianfan Platform helps industries build the underlying platform and provides handy tools to help enhance the ability of enterprises to use large models, which accelerates the large-scale implementation of large models. At the same time, through the layout and implementation of AI large models, and through in-depth

cooperation with the industry through the Qianfan large model platform, Baidu itself can also go through the “macrocycle" of development by releasing "intelligent productivity.”