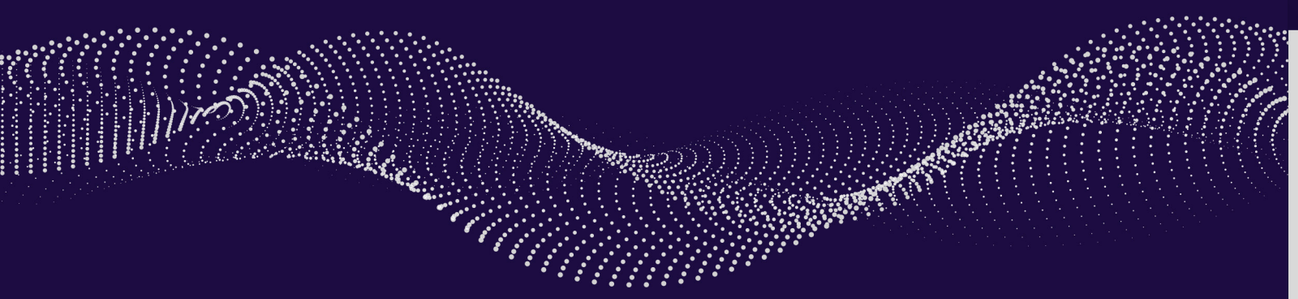
Alibaba Cloud

www.alibabacloud.com



Alibaba Cloud

AI FORWARD

Enabling Smarter Business Strategies and Pioneering Solutions

Alibaba Cloud

www.alibabacloud.com

Contents

Foreword

As we step into the promising horizons of 2024, it's with immense gratitude and enthusiasm that we bid adieu to a year that encapsulated extraordinary growth and transformation. The pace at which technology advanced throughout 2023 has been nothing short of astonishing and I am deeply proud and appreciative of our collective achievements in 2023.

The year was marked by momentous shifts in technological landscapes and market dynamics. Even consumer media was dominated by talks of technology and the strides industries have taken in 2023. Reverberating with the buzzwords "generative AI" and "large language models," 2023 will be remembered as the year that catalyzed the next technological revolution.

Generative AI and LLMs triggered a global change and bestowed businesses with a surge in computing power, igniting the next wave of industry-wide transformation. New AI technologies that took center stage in 2023 are redefining the boundaries of what technology can achieve, and we are yet to see all its applications. The initial half of 2023 witnessed enterprises gearing up to embrace generative AI's transformative capabilities, prompting a boost in investments toward AI-related research and applications. We are confident that these investments in AI technology will make a lasting impact in the coming years and decades. Most of these innovations were driven by hubs in China and Southeast Asia, drawing diverse entities and businesses into the fold of a high-potential AI ecosystem.

Alibaba Cloud has always placed itself as a pioneer of technological innovation. In 2023, we continued this trend, positioning ourselves as a pivotal force in the AI space. One of our key pillars, AI has been driving innovation at Alibaba Group since its earliest days. Looking back on those years, one can see the groundbreaking strides Alibaba Cloud took to advance AI and data technologies for Alibaba Group and the world. Our AI journey began with a technological core that enabled the scaling of clusters to encompass over 10 thousand servers and empowered the processing of vast data volumes on the ODPS platform.

In 2023, Alibaba Cloud influenced and witnessed key breakthroughs in AI technology. The unveiling of the Tongyi Family reiterated our commitment to advancing large language model research and solidifying our stance as frontrunners in language model development within China. We also made bold moves to further AI development, opening up our pre-trained large language models, including Qwen-72B and Qwen-1.8B to the open-source ecosystem.

This pivotal step aligns with our vision of an open and inclusive cloud platform in the intelligence era. Our open-source Tongyi LLM is the leader on the Open LLM leaderboard on HuggingFace among pre-trained models - a noteworthy recognition of our vision. Furthermore, our initiative of the "Model-as-a-Service" concept envisions a future characterized by an inclusive AI ecosystem flourishing on robust cloud infrastructure, conducive platforms, and an expansive spectrum of AI technologies and models.

In 2024, we are on the path to democratizing AI, as stated by our executives Eddie Wu, CEO of Alibaba Cloud and Alibaba Group, and Joe Tsai, Chairman of Alibaba Group. We also envision a 'third wave' of digital transformation powered by LLMs and AI technologies.

The cloud will need to enable development and innovations in AI, and cloud computing is the basic prerequisite for AI innovation. As a leader in the cloud computing industry, it is a trend that we are already experiencing. In China, over 80% of all technology companies and large model-related businesses leverage Alibaba Cloud's various cloud-based services and solutions for AI development.

The ultimate innovations in AI will come from various businesses in the market that will leverage their industry-specific expertise, use cases, and understanding of customer needs. They will leverage their knowledge of business use cases to improve productivity and user experiences. Furthermore, businesses will also need to incorporate AI into their products and services. For many, leveraging AI is a necessary business practice to gain a competitive edge and keep up with the trends. No business wants to lag for "not using AI".

AI will serve us as a powerful tool in 2024 and beyond but it is important to remember that innovation is the true driving force for progress. Innovation will unlock what AI is truly capable of and shape the technologies that will grace us and redefine the frontiers of technology in 2024. I'm excited that we're going to kick start a promising year with a wonderful event: the AI and Big Data Summit in Singapore.

I'm looking forward to the opportunity to spark more inspiration and insights from our partners, customers, industry leaders, and technology experts, and together, we'll gear up for an amazing 2024. As we embark upon the uncharted territories of 2024, we invite you to join us in this journey of exploration and transformation, as together, we shape the future of technology.

Selina Yuan
Vice President of Alibaba Group;
President of Alibaba Cloud
Intelligence International

NEXT

Alibaba Cloud is taking groundbreaking strides for Gen AI, Big Data, and other related technologies in pursuit of revolutionizing the technological landscape. Our NEXT CLOUD embodies our technological excellence and amplifies the transformative potential of AI by pushing the boundaries of computing capabilities.

By integrating these cutting-edge technologies, Alibaba Cloud aims to foster unprecedented advancements, such as enhanced machine learning algorithms, AI-powered applications, and next-gen data analysis. Alibaba Cloud is committed to shaping the future of AI, ushering in a new era where intelligent AI solutions converge to redefine what is possible for all industries.

02 HOW LARGE LANGUAGE MODELS STEER THE COURSE OF CLOUD COMPUTING IN THE AI ERA

Model development and training heavily depend on the infrastructure support provided by cloud technology. Simultaneously, cultivating a thriving ecosystem involves establishing a comprehensive model ecology, from foundational to industry-specific models. Alibaba Cloud is dedicated to continuously enhancing the entire application development paradigm, aiming to streamline the integration of cloud technology and AI.

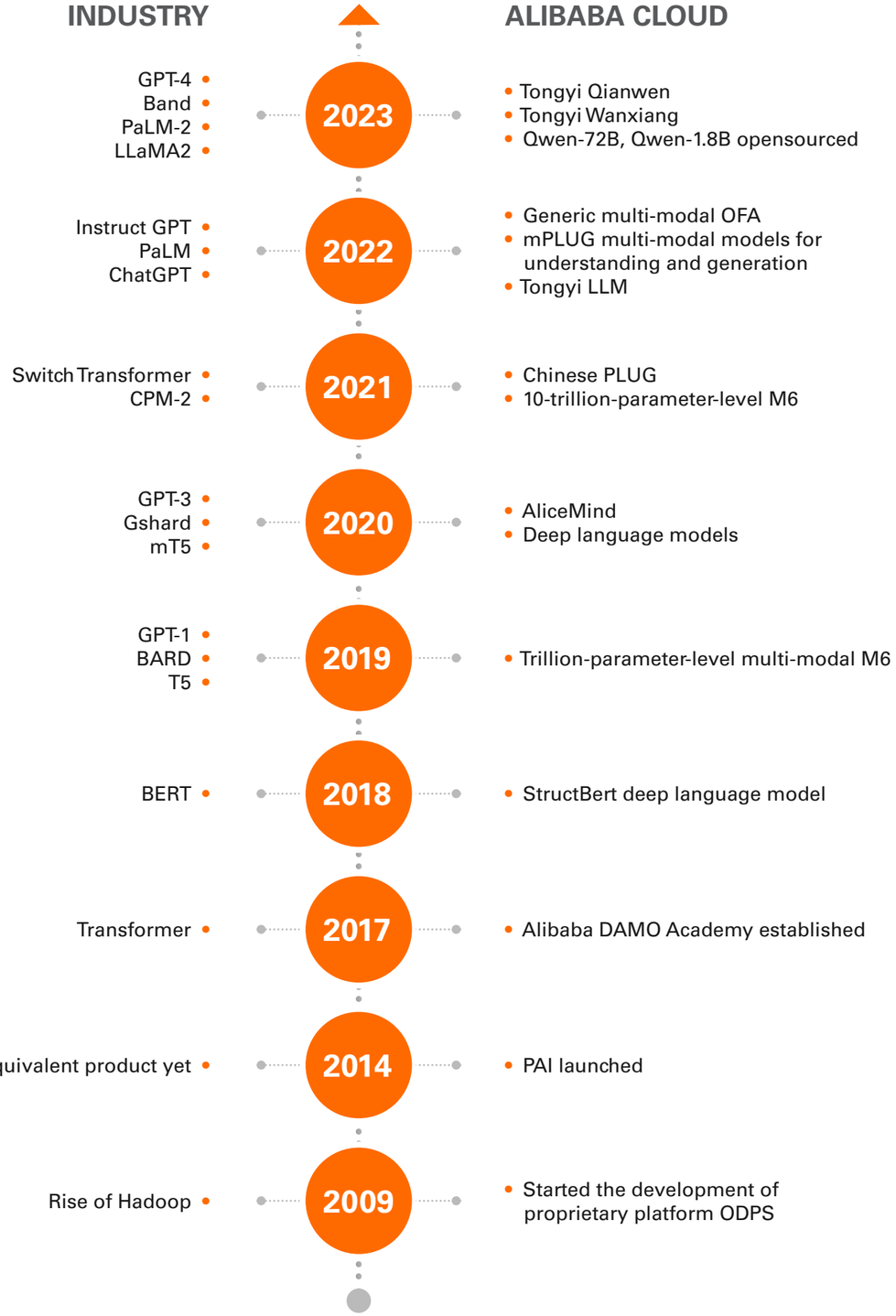
Jingren Zhou
Chief Technology Officer, Alibaba Cloud

Alibaba Cloud: Frontier of Innovation with Decades of AI and Data Technology Advancements

The technological and service capabilities that Alibaba Cloud has built for AI Large Language Models (LLMs) today are not merely the result of recent AI research initiatives. Instead, they are firmly rooted in the decades-long development, accumulation, and evolution of Alibaba Cloud's technology.

In 2009, Alibaba Cloud initiated research and development of big data computing services with ODPS, thereby establishing a leading technological capability in large-scale data processing, leveraging computing at scale. This advancement paved the way for the launch of PAI (Platform for AI) in 2016. Later, Alibaba Cloud released the Tongyi Large Language Model at the AI Conference in September 2022 and introduced the concept of "Model as a Service" (MaaS) in the industry at the Apsara Conference in November 2022.

Alibaba Cloud's AI Progression



We believe that MaaS provides a unique direction for the future development of AI and, when implemented effectively, it will significantly enhance the growth of the AI ecosystem. On the other hand, MaaS depends on infrastructure, particularly necessitating deep integration with cloud computing infrastructure, which creates opportunities for new AI tools.

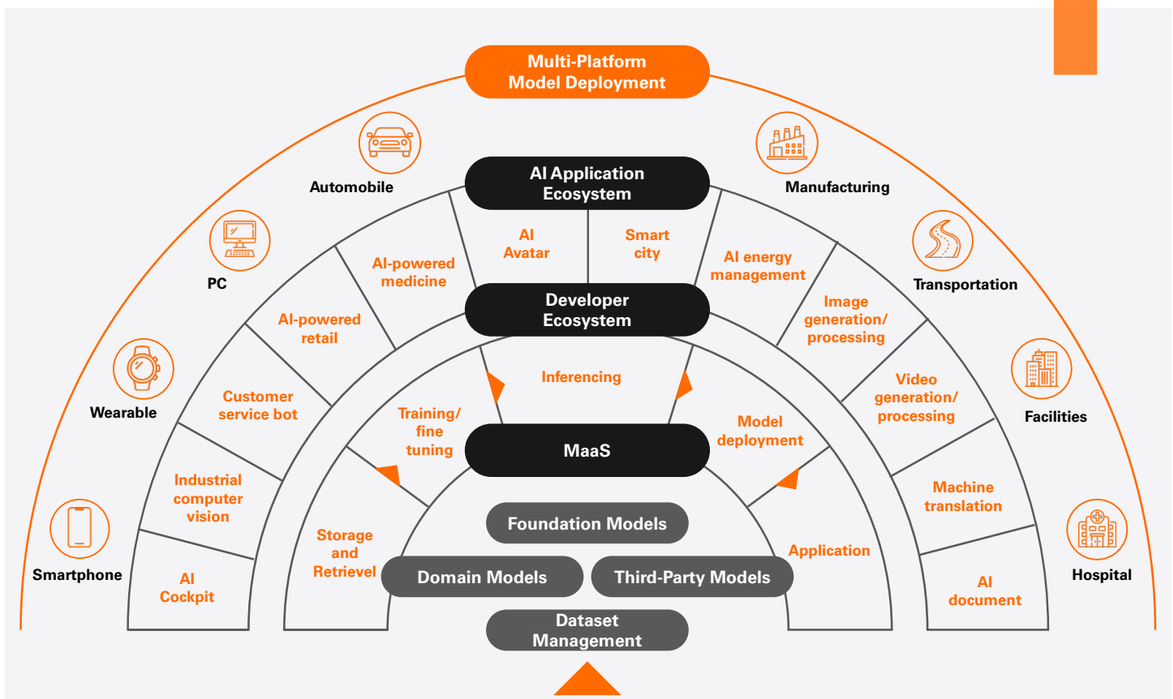
The development of cloud computing will certainly continue to accelerate in the future, with cloud-based businesses and enterprises speeding up their innovations on the cloud. With the deep integration of cloud technology and intelligence, AI will be ubiquitous. Founded on cloud technology infrastructure and driven by AI as the engine, Alibaba Cloud will also bolster innovation in the intelligent era, laying a solid foundation for the rapid development of AI across all industries.

Revolutionizing AI with Cloud as the Foundation and AI Models as the Core

To accelerate breakthroughs in AI application challenges, we believe that the first step is to build a service system centered around AI models.

The concept of Model as a Service (MaaS) implies that at the fundamental layer, models have become crucial elements of production. Starting with model development, which encompasses tasks such as data cleaning, feature engineering, model training and tuning, and model services, we must design products and technologies aligned with the entire lifecycle of models.

Furthermore, another layer of significance lies in directing focus on the ecosystem development surrounding AI models. This involves a focused approach to industries and scenarios, promoting the continuous development and innovation of layered model structures. The LLMs ecosystem is already evolving, particularly the pre-trained LLMs tailored for multimodal scenarios, marking a prevailing trend. Their "large-language" and "pre-trained" attributes enable their capabilities to generalize easily, making them the foundation for upper-layer application technologies.



These models can effectively support the deployment of numerous AI applications, addressing challenges and complexities involved in traditional AI applications. As a result, this significantly accelerates the application of AI technology across a multitude of industries.

With the Model as a Service (MaaS), Alibaba Cloud bases its approach on foundational LLMs to develop domain-specific models and adapt to multi-end model service deployments. Within this approach, Alibaba Cloud offers customers a comprehensive model service through its open-sourced model platform, covering the entire lifecycle management of numerous LLMs essential for development. This approach integrates seamlessly with the diverse application ecosystems, enabling the deployment and application of industry or enterprise-specific models.

Shaping Tomorrow: Bringing AI Models into the Industry

Alibaba Group embarked on its journey into multimodal pre-trained Large Language Models (LLMs) in 2019, aiming to integrate knowledge to effectively accomplish the recognition and understanding of all subjects.

In 2021, Alibaba Group released a Chinese pre-trained multimodal LLM initially comprising tens of billions of parameters. Subsequently, the parameter count increased from tens of billions to trillions and eventually to hundreds of trillions, accompanied by active involvement in the research and development of related multimodal models.

- In September 2022, all these models were brought together to form the "Tongyi" series of LLMs, providing a unified foundation for Alibaba Cloud's multimodal models. Beyond providing a unified training framework, the significance lies in fostering collaboration and open-sourcing. Industries can leverage this platform to create and develop secondary models and progressively build a comprehensive model ecosystem. The "Tongyi" platform already features a hierarchical and modular structure with industry models. Industry models can be customized on top of the pre-trained models to address over 200 different business scenarios currently.
- April 2023 marked the introduction of "Tongyi Qianwen," an ultra LLM capable of understanding human instructions and engaging in multi-turn interactions with humans. It also integrates multimodal knowledge comprehension and can connect with various externally enhanced APIs.
- Expanding the Tongyi family, on June 1st, "Tongyi Tingwu," an AI product focusing on audio and video, was officially launched, becoming the first LLM product in China for public beta testing, integrating advanced voice semantics, multimodal algorithms, and more.
- On July 7, during the 2023 World Artificial Intelligence Conference, "Tongyi Wanxiang" designed for AI painting creation started targeted invitation testing. This model serves as a valuable aid for humans in image creation and holds the potential for diverse applications in areas such as art design, e-commerce, gaming, and cultural creative scenarios.

Moving ahead, Alibaba Cloud plans to provide a comprehensive model library across diverse industries. Crucially, we aspire to unlock the potential of these models through collaboration with our partners, facilitating further development, fine-tuning, and integration with industry-specific knowledge for practical applications. This initiative aims to offer a wide array of flexible invocation scenarios, contributing to the comprehensive and intelligent transformation of business systems.

03 ACCELERATING INNOVATION WITH LLM-FRIENDLY AI AND BIG DATA PLATFORMS

The emergence of large language models (LLMs) and the Model-as-a-Service (MaaS) delivery model defines a new standard for AI development. As the demand for enormous computational power continues to grow, it presents unique challenges to the very foundations of AI engineering. In an era when businesses are integrating large language models (LLMs) into their processes, we need highly efficient AI + Big Data platforms that incorporate computing, development, and processing capabilities to ensure the support for business-wide innovation in the AI era.

Junhua Wang
VP at Alibaba Cloud & Head of Computing Platform

During the 2023 Apsara Conference held on October 31st, 2023, we announced enhancements and upgrades to Alibaba Cloud's AI and Big Data platforms. These upgrades were specifically tailored to meet the surging demand for Large Language Models (LLMs) in various business landscapes.

Supporting Varied Business Demands with Multimodal and Flexible AI Development

In the fast-evolving AI sector, there's a growing need for specialized skills among AI developers and flexible solutions to facilitate AI development. To address this, our AI platform, PAI 4.0, simplifies LLM-based AI development by offering a comprehensive and user-friendly support system that reduces barriers to entry.

Whether it is developers in the deep learning community tasked with defining model structures and development processes, groups with massive large-scale computational tasks, or individuals who need to efficiently and quickly link training and inference tasks - all can carry out R&D through PAI. The platform accommodates and supports popular frameworks, open-source models, and deployment scenarios, serving as a one-stop development solution.

Enhancing LLM Performance and Enabling Deeper Big Data/ AI Integration with Efficient Data Services

Throughout the machine learning development stage, 80% of R&D efforts are devoted to data preparation and optimizing data quality for improved large language model performance. Consequently, there is an increased emphasis on data processing and analysis. Alibaba Cloud offers comprehensive Big Data solutions for data accumulation, cleaning, modeling, computing, and services to streamline this preparation phase.

Big data is now an integral part of AI development. To accommodate business requirements for Big Data, Alibaba Cloud's MaxCompute has enhanced its DataFrame capabilities and released MaxFrame, a distributed computing framework 100% compatible with data processing interfaces like Pandas. With just one line of code, native Pandas can be converted to MaxFrame. This innovative approach bridges the gap between data management, large-scale data analysis, processing, and ML development. By eliminating barriers between Big Data and AI, this advancement improves efficiency and enables better development cycles.

Current generation AI driven by large language models, demands prompt data availability. Alibaba Cloud's Flink+Paimon solution provides a one-stop solution for data ingestion, real-time processing, and analysis within a data lake, bolstering Flink's real-time computational strengths for AI applications.

Additionally, vector capabilities are updated in Hologres, OpenSearch, and Elasticsearch, catering to performance needs across various use cases. The release of DataWorks Copilot combines big data platform features with large model AI capabilities, promoting the fusion of AI with business processes to generate new value.

Following the upgrades in big data AI products for the era of large language models, Alibaba Cloud's big data AI products have been fully transformed to Serverless architecture, providing customers with ready-to-use, cost-effective products using a pay-as-you-go model. As the foundational infrastructure for AI in the large language model era, Alibaba Cloud's AI + Big Data platform will persist in investing resources to support diverse industry advancements.

NEO

With a focus on robust infrastructure and industry-specific expertise in retail, finance, manufacturing and more, Alibaba Cloud pioneers global digital transformation journeys for millions of customers. The NEO CLOUD represents our solutions that help customers globally optimize their operations, enhance their efficiency, and increase their competitiveness.

By fusing the latest cloud technologies and breakthroughs in AI for our customers we aim to introduce AI-powered transformation journeys in 2024 and beyond. With our breakthroughs in AI and their applications in diverse business scenarios, we aim to exemplify the profound influence AI will wield in advancing world-class projects.

04 THE BORDERLESS AI ECOSYSTEM: HOW ALIBABA CLOUD IS BUILDING THE FOUNDATIONS FOR AI INNOVATION

Foundation models have revolutionized AI with the ability to comprehend and generate human-like output. Alibaba Cloud was among the first companies in China to explore the potential of such large foundation models and by 2020, we concluded that comprehensive multimodal models are the next step toward achieving human-like intelligence with AI. AI agents developed with multiple foundation models and reinforced by learning algorithms possess the potential for advanced learning, reasoning, and real-world interaction.

Guo Dongliang
Vice President, Product and Solution, Alibaba Cloud International

Intelligent AI agents could transform AI's passive support role, potentially acquiring superhuman competency to solve complex problems and automate decision-making. Yet, to achieve truly intelligent AI agents, progress in multimodal foundation models is essential for improved accessibility and output quality. We believe that today's challenge lies in integrating various AI modalities into a singular model capable of solving real-world problems. For example, combining visual content capabilities into a language model to generate graphics with text copy from a single prompt.

Since humans gather knowledge through vision, language, hearing, and more, these elements cannot be treated as independent entities. The human brain seamlessly connects these inputs into a unified knowledge system, and the most advanced AI needs to similarly combine data from all sources for centralized processing. We believe multimodal models will inevitably merge information from any source into a complex space, enabling unified processing that combines image/video, text, and audio.

Building the Bedrock for Multimodal Models with Alibaba Cloud Tongyi

With the introduction of our "Product Model" approach to AI foundation models, as opposed to the industry-standard "Industry Model," we are not just innovating; we are redefining the way AI can serve businesses and individuals globally. The Tongyi model is a testament to our commitment to providing an end-to-end AI ecosystem for multimodal AI development. It is designed to be the bedrock upon which developers and AI startups can build, innovate, and deploy AI applications across various industries.

Our eight product models serve as showcases for the potential of what large AI models can achieve. They also act as a bridge between Alibaba Cloud's technological prowess and the practical needs of businesses seeking to leverage AI to its fullest. In line with our vision to be "the most open cloud," we have opened up our APIs for these application models. Here's how we're fostering a collaborative community for developers and enterprises—ranging from industry integrators to AI startups and independent AI developers.

- Enhancing accessibility for Tongyi by collaborating with 60+ leading industry integrators and independent software vendors as of October 2023
- Implementing our Tongyi foundation models in various fields such as office automation, cultural tourism, electric power, governance, healthcare, transportation, manufacturing, finance, and software development
- Strongly supporting third-party development of large foundation models and sharing technological advancements with developers
- Partnering with AI service providers such as IDEA Institute, Baichuan Intelligence, Zhip AI, and Shanghai AI Lab through the Modelscope community
- Empowering developers with free GPU computing power for experimenting with large foundation models, having already provided over 30 million hours of computation

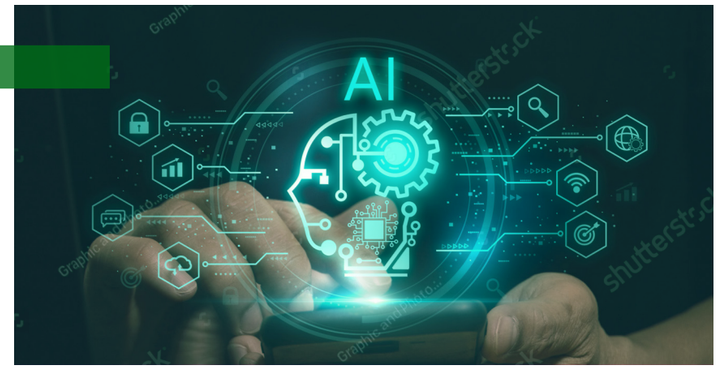
Smart Agents to Assist Enterprises in Foundation Model Adoption

Additionally, we want to ensure that the Tongyi models act as secure and compliant foundation models for enterprises. A trend we've noticed is that enterprises actively pursuing the adoption of large models face significant challenges, including inaccuracies in model outputs, cybersecurity risks, and potential compliance violations. We expect a rise in smart agent-based solutions to assist enterprises in navigating the adoption journey of large models.

For instance, in 2023, Alibaba launched Tongyi Bailian, a comprehensive platform prioritizing security, customization for high-quality model output, and user-friendliness through smart agents. This platform guarantees data security by housing enterprise data in private cloud environments (VPC) and offers a powerful semantic model, General Text Embeddings, for accurate document retrieval and enhanced output quality.

Enterprises also gain access to a robust toolkit, including vector databases and a process flow designer for application development, as well as data annotation and processing services to train top-notch models. Moreover, it provides thorough model evaluation services and a multi-layered security approach, incorporating content filtering and safety mechanisms to prevent harmful content and ensure compliance with regulations. This makes it a secure and effective solution for enterprises leveraging the potential of large models.

The Tongyi product model strategy is more than a set of services; it's a commitment to an open-source AI community. It's a journey we are on together with our partners, collaborators, and the general public to push the boundaries of what's possible with AI. Join us as we continue to democratize AI, making it accessible, versatile, and open for all, forging the smarter future we all aspire to see.



05 MAKING GLOBAL BUSINESSES SIMPLE WITH AI EXCELLENCE

Based on an interview with Kaifu Zhang from Alibaba International Digital Commerce.

Kaifu Zhang
Head of AI Business, Alibaba International Digital Commerce

Q. Many people believe that the Gen-AI Boom is just a tool for improved productivity. How do you interpret the transformations we have witnessed in 2023?

AI is reshaping the world in two key ways. Firstly, it serves as a valuable tool to enhance productivity and foster cost-efficiency—an application widely embraced today due to a consensus on its efficacy. The prevalence of AI adoption within Alibaba Group's business ecosystem is also evident in various facets, such as AI-driven marketing, AI content generation, and AI-powered customer services.

AI also brings another dimension to the perception of technology. It's not just about spending less but about accomplishing more. For example, the Alibaba team tried using AI to enhance its online marketing. They realized that personalizing product recommendations for consumers has become highly nuanced and tailoring suggestions to each user is a time-consuming process. However, constructing individualized marketing copies manually for every user proves unfeasible due to the complexity and volume of the task.

However, generative AI has now made that possible. The most significant benefit of this change is saving the overall cost of marketing content production, providing higher efficiency. Hiring the right people poses a considerable challenge for smaller enterprises. Imagine the difficulty of hiring a qualified visual designer with solid skill sets in their specialization and proficiency in multiple languages. Picture the challenge of finding an attorney specializing in compliance matters within the finance industry, capable of effectively interfacing with financial organizations. Filling such high-quality and low-cost requirements can be a massive challenge for a small enterprise, but generative AI can help navigate this by filling in the gaps. With AI, businesses can not just save costs but do things in a new way, transitioning from the impossible to the possible.

Q. What do you perceive as the biggest differentiator in the latest Gen-AI advancement? Is it just a "better" AI tool?

Zhang believes that the evolution from translation to localization is what defines Gen-AI advancement. According to his experiences in cross-border e-commerce, the language barrier remains a substantial impediment to cross-cultural communication. He gave an example of a supply chain owner based in China struggling to talk with and trying to convince interested global buyers in English while his customers' native languages could be French, Spanish, German, or Arabic. Traditional "translation" methods sometimes fall short of achieving the desired communication outcomes. It's crucial to present product offerings and key benefits to customers in a manner that resonates with their own language and culture.

The Large Language Models (LLMs) present a new translation approach. An LLM-based translator can virtually "understand" and based on that understanding, "create" an expression for the target audience. Additionally, it has the capability to articulate perfectly in the target language. With LLMs, businesses can truly achieve "localization." Here's an example, say an online shop owner has developed a copy for a product they are marketing. When posting it on online marketplaces, they realize their audiences may speak a different language. Let's consider that the potential customers are Spanish-speaking and from Mexico. They can ask the AI-based writer to take the copy and create a similar copy for a Spanish-speaking audience based in Mexico. The AI is actually recreating the message for them instead of doing a word-for-word translation; it involves adapting the message to resonate better with the audience. For major languages like English, Spanish, and French, the evolving linguistic and cultural landscape has created many linguistic variants among different speaking populations. He further elaborated on how expressions for the same concept can vary widely across countries and cultures. For example, what do you call carbonated beverages in your hometown? Is it called soda or pop? Do you use the metric or royal system for measurement in your country? LLMs can account for these cultural and linguistic landscapes, making their suggestions more effective.

Overall, the advancement in generative AI and current LLMs is two-faceted. Firstly, there is a significant breakthrough in AI's language proficiency. Secondly, AI can now incorporate other factors, such as cultures and customs, seamlessly integrating into diverse vocabularies and "schemes of things."

Q. In the context of the retail business, do you think there are more use cases?

In online commerce, marketing is a daunting and expensive task due to its content-intensive nature. For example, how you write emails - considering elements such as the subject line, keywords, benefits highlighted in the main body, and other content will significantly affect how your electronic direct mail (EDM) performs when viewed by the recipient. For an online shop owner aiming to send a marketing email to 100 thousand recipients, optimizing content becomes a crucial consideration to appeal to the majority of potential customers. While a hired content writer may only produce a few email versions daily, AI offers the capacity to generate personalized copies for each recipient, contemplating what might appeal to specific users and what words might trigger the buying decision. It's worth noting that the capabilities of AI extend beyond textual creation; it can also generate visual content, providing a comprehensive toolset for online marketing endeavors.

For example, Aidge is an industry-specific platform for empowering enterprises in the cross-border commerce business, integrating know-how, use cases, and optimizations made for the domain. With Aidge, customers can unlock improved efficiency and cost savings, effective content localization, and content creation functionalities.

Launched by Alibaba International, Aidge is an AI platform that's on a mission to simplify international business through AI. Powered by in-house LLM and AIGC models, and enriched with global e-commerce operation experience and insights, Aidge provides a comprehensive AI suite to help businesses of any size break down language and cultural barriers, compensate for talent shortages, and focus on business results.

Q. How do you envision the future landscape of artificial intelligence unfolding?

According to Zhang, the AI landscape is not flat but rather an ecosystem comprising diverse terrains. It's not feasible or beneficial for every business to train its own large models, dealing with the engineering complexities of managing and utilizing these models. Opting for a bigger approach isn't necessarily the golden strategy. Going forward, the Large Language Model (LLM) domain is evolving into a diverse ecosystem with various forms of coexistence. Alibaba Cloud is already taking steps to build the AI ecosystem that businesses need.

Within this ecosystem, there will be pioneers pre-training and refining the large-scale models, industry leaders working on sector-specific models based on fine-tuning, industry-specific know-how, and data resources, along with a proliferation of applications discovering their optimal use cases in this dynamic landscape. Alibaba Cloud's IaaS-PaaS-MaaS vision is also a manifestation of the future-forward AI ecosystem.

Q. Who will be the key player in this wave of AI revolution? And what's your opinion about the concept of "agent?"

Real-world AI use cases have become more valuable in this new AI revolution. As more players in the market flock towards this new trend and try to redefine their existing functions with AI, they're facing the ultimate question, "How many functions can they leverage and empower?" It's down to how many APIs they can open up from their business side. So, the question, "What can AI do for us?" is no longer relevant. The question businesses must ask now is "What can we do with AI?" Our functions and use cases will allow AI to make decisions and take action on our behalf.

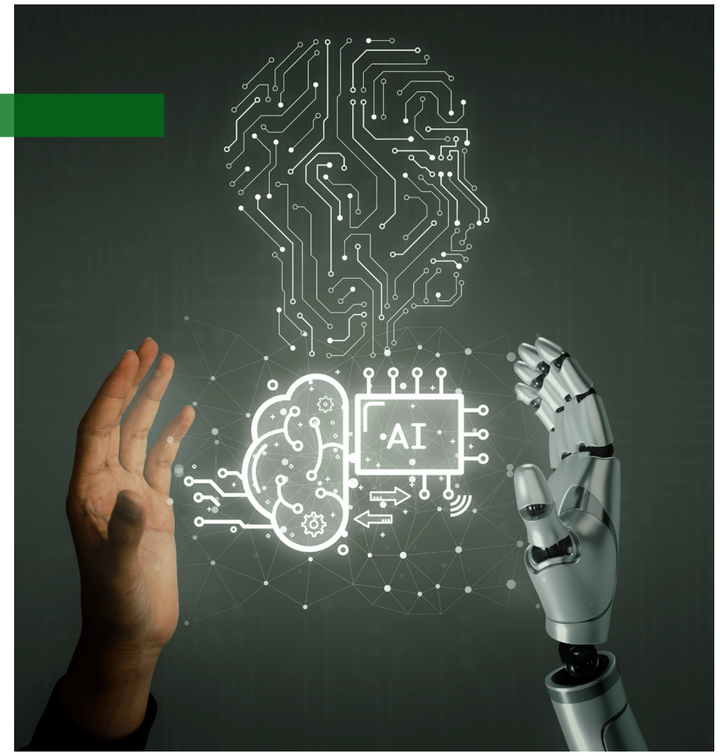
Over the past year, the concept of an "agent" has emerged as a prominent trend in the evolving AI landscape. However, according to Zhang, this concept has been over-generalized. As for agent engineering platforms, the top players have already paved the way for all. Looking ahead, there are two potential trends in the realm of agents. On one side, domain-specific agents may flourish, capitalizing on their domain expertise and tailored use cases. On the other side, agent-centric ecosystems could be led by industry giants leveraging their extensive "reach" and substantial "user base."

Q. Many people would implement AIGC solutions in the retail business just for faster and cheaper image production. Do you agree with this claim?

A question everyone must always ask themselves is whether AI will actually drive cost savings for society in the long run. Surveying the value chain across design, production, and marketing, people may argue that AI's contribution will eventually be perceived as minor cost savings in the form of increased productivity throughout the lifecycle. For instance, consider AI-generated images utilized for product detail pages in the apparel business. However, what may elude most people's perception is that the product image serves a distinct role, contributing value that extends beyond conventional measurements.

Taking a product image as just an interface between the consumer and the product, the value that AI can deliver is merely the cost savings achieved by eliminating the need for a human visual designer. Of course, the image can now be delivered within a shorter turnaround time with the assistance of AI.

However, you may automatically take the image for an "endpoint," whereas it could be a starting point. Consider it as the tail of a dog, but in reality, it's the tail that wags the dog. Instead of presenting a finished product, a supplier can showcase twenty images representing diverse but promising designs to its target consumer group. After consumers express their preferences by placing orders, agile production and supply chains can fulfill these orders promptly. Testing the "prototypes" with real-world customers allows the supplier to determine the best candidates for its next bestseller. With the help of AI, more design prototypes and variations can be tested. Imagine how it will totally redefine the online apparel sales business. Leveraging the sales performance data, AI can make better-informed designs incorporating the design elements most likely to appeal to the target consumer group.



Q. Which side do you take in the "blessing or curse" debate?

Since its inception, AI has been a fiercely debated technology concept. The debate usually revolves around two primary arguments. The first is centered on the maturity of the technology. For example, one of the key challenges that impede universal adoption is "hallucinations." In the long run, the evolution of technology will address this challenge. AI will mature over the next few years or even decades as the data scale and model sizes continue to grow. In the short run, engineering approaches like retrieval augmented generation (RAG) can address "hallucinations" in predefined use cases while allowing businesses to leverage the productivity increase offered by generative AI. We try to get the AI bot to base its answers on facts that can be sourced from the database. Another frequently raised argument questions whether AI just saves some costs by displacing human effort. Zhang contends that the most significant value of using AI will be fulfilled with new possibilities being discovered, not only existing functions being eliminated with automation.

06 TAKING FASHION FORWARD WITH TECH: AN EXCLUSIVE INTERVIEW WITH HAN NENG WONG FROM CHARLES & KEITH

Established in 1996, CHARLES & KEITH is a Singaporean fashion label that creates collections (shoes, bags, eyewear, accessories, and costume jewelry) to enhance life and work. From a single footwear shop in the suburbs to an international network offering memorable retail experiences in physical stores and online, CHARLES & KEITH's humble beginnings are a testament to its challenger spirit, which has put Singapore on the map as a place of originality and creativity. To understand how a fashion brand converges with emerging technologies, we interviewed Han Neng Wong, Senior Manager, IT & Cyber Security at CHARLES & KEITH, who shared their insights into how they leverage AI & Alibaba Cloud technologies.

Han Neng Wong
Senior Manager, IT & Cyber Security
CHARLES & KEITH GROUP

Q. As we know, CHARLES & KEITH is a popular brand with innovation and creativity in its DNA. Do you think the beauty and fashion business needs to inject technology to deliver better customer satisfaction?

In the ever-evolving fashion industry landscape, integrating technology is critical for delivering better customer satisfaction. Technology enhances operational efficiency and plays an important role in providing a seamless and engaging customer experience.

The integration of technology in our business strategy is geared towards improving the overall customer experience. This involves personalized interactions through online platforms and virtual try-ons, allowing us to understand and cater to individual preferences. By leveraging technology in supply chain management optimizes inventory, reduces lead times, and ensures prompt fulfillment of customer demands, contributing to a positive brand image.

We leverage analytics to gain insights into market trends and customer behaviors. This enables us to fine-tune our product offerings and optimize our marketing strategies specifically tailored to our target audience. For this scenario, we deploy Alibaba Cloud MaxCompute to enhance our big data capabilities in our digital transformation journey. Finally, our emphasis on omnichannel integration creates a seamless transition between physical stores and online platforms, providing customers with a consistent brand experience across various shopping channels.

Injecting technology into our operations is not just beneficial; it is essential for staying competitive in the dynamic world of fashion. It enables us to adapt to changing customer expectations, foster innovation, and ultimately deliver a superior and memorable customer satisfaction experience.

Q. How does CHARLES & KEITH embrace the technological revolution to fuse fashion and technology?

At CHARLES & KEITH, we have adopted several strategies, leveraging technologies to enhance our fashion brand. Here's what we do to fuse fashion & tech:

- **Data Analytics for Personalization** - We leverage big data analytics to gather insights into customer preferences, shopping behaviors, and trends. With our analytics, we implement personalized marketing strategies, recommending products based on individual preferences and tailoring promotions to specific customer segments. To get the most value out of our data, we leverage Alibaba Cloud Open Data Platform and their cutting-edge MaxCompute & Hologres.

- **Smart Retail Solutions** - We recently introduced smart retail solutions based in Alibaba Cloud for our physical stores, such as interactive displays and smart mirrors at selected outlets to enhance the in-store experience, providing customers with additional information about products, styling tips, and promotions.

- **Sustainability and Technology** - CHARLES & KEITH embraces sustainable practices by incorporating technology to reduce the environmental impact of the fashion industry. We achieve this by exploring and implementing innovative materials, production processes, and supply chain technologies that align with CHARLES & KEITH's commitment to sustainability.

- **Culture Developed Around Innovation** - We encourage a culture of continuous innovation within the organization. We support our employees in exploring and implementing new technologies, staying updated on industry trends, and participating in training programs to enhance their technological skills.

Q. What role does online presence play in CHARLES & KEITH's business?

The online presence of CHARLES & KEITH plays a pivotal role in various aspects of the business, contributing to its overall success and growth. Here are key roles that our online presence serves for CHARLES & KEITH.

- **Adaptability to Market Trends** - The online platform allows CHARLES & KEITH to quickly adapt to changing market trends. Whether responding to emerging fashion preferences or incorporating new technologies, the brand can stay agile and relevant, catering to the dynamic nature of the fashion industry.

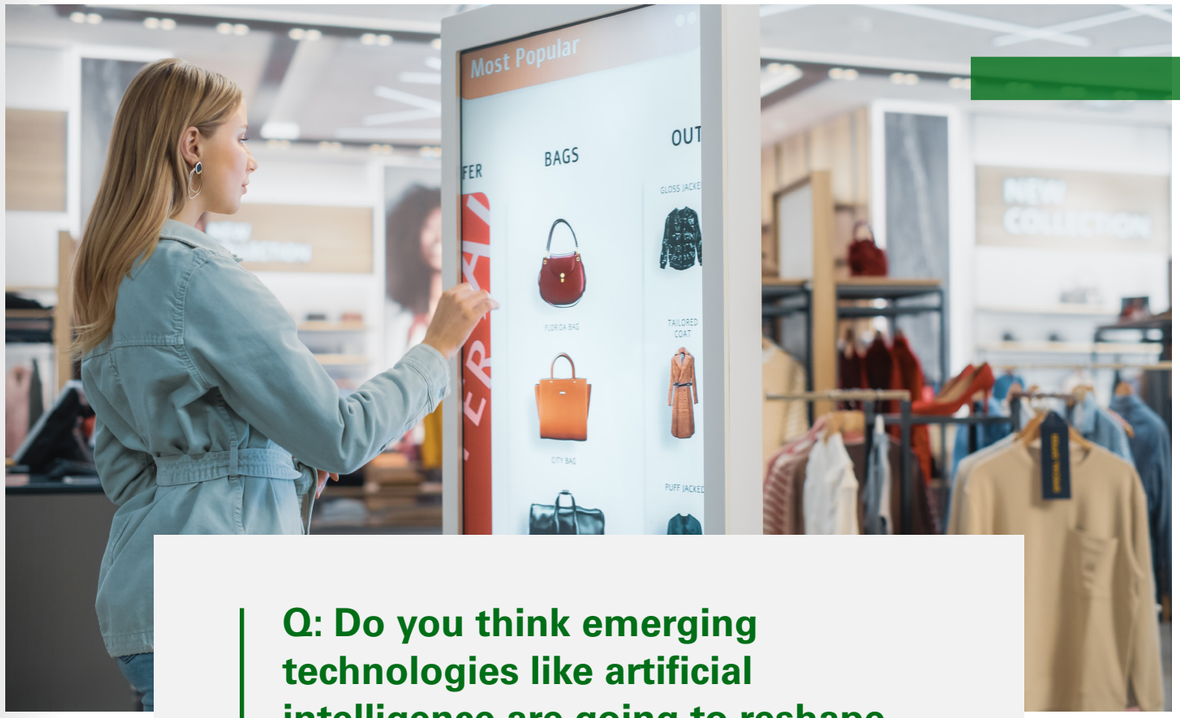
- **Global Reach and Accessibility** - Our online presence allows the brand to transcend geographical boundaries and reach a global audience. Customers from different parts of the world can access the latest collections and trends and make purchases, expanding the brand's market reach beyond physical store locations.

- **Omnichannel Integration** - It serves as a key component in achieving omnichannel integration. Customers can seamlessly transition between online and offline channels, enjoying a cohesive brand experience. Features like online ordering with in-store pickup or returns contribute to a unified customer journey.

Q. What do you see as the most critical success factor in leveraging data intelligence and AI effectively? Is it data, algorithms, infrastructure, or something else?

At CHARLES & KEITH, we've learned that to leverage data intelligence and AI effectively, all the elements - data, algorithms, and infrastructure are crucial, and their synergy is necessary for success. However, the most critical success factor often lies in effectively managing and utilizing data.

- **Quality Data** - The foundation lies in clean, accurate, and relevant data, which is essential for training robust machine learning models and informed decision-making.
- **Data Governance** - Establishing robust data governance practices is crucial. Clear data ownership, security protocols, and compliance measures ensure responsible and ethical data handling, maintaining trust with customers and stakeholders.
- **Data Integration** - Gaining a holistic view of the business requires seamless integration of data from various sources, encompassing customer behavior, market trends, and internal operations.
- **Data Accessibility** - Ensuring relevant stakeholders have timely access to the right data is key. Efficient data storage solutions, access controls, and user-friendly interfaces democratize data access, empowering teams throughout the organization.
- **Advanced Analytics Algorithms** - Beyond foundational data, advanced analytics algorithms and staying updated on AI advancements are vital. Investing in these technologies ensures meaningful insights and predictions are derived from the data.
- **Scalable Infrastructure** - Alibaba Cloud provides us with a flexible and scalable infrastructure for handling increased data loads and adapting to evolving business needs.
- **Talent and Skill Sets** - Critical success lies in building and maintaining a skilled team in data science, machine learning, and AI. The talent we hire is responsible for interpreting data, developing algorithms, and optimizing AI models to ensure that CHARLES & KEITH's data generates maximum value.



Q: Do you think emerging technologies like artificial intelligence are going to reshape your industry entirely or just make things easier with small but impactful changes? Do you have some examples or practices to share?

Artificial Intelligence has the potential to significantly reshape the fashion industry, bringing about both transformative changes and incremental improvements.

A few changes and improvements that AI will bring to the industry include:

- **Predictive Analytics for Fashion Trends**
  AI can analyze vast amounts of data from social media, fashion blogs, and other sources to predict upcoming trends. This insight can guide design decisions, helping fashion companies like CHARLES & KEITH stay ahead of the curve and create products that resonate with consumer preferences.
- **Employee Productivity**
  AI tools, such as automation software, can streamline various internal processes, freeing up employees to focus on more strategic and high-value tasks. This can lead to increased productivity and efficiency within the organization.

Q. What's your take on the latest developments in generative AI? Will it make a significant impact on industries, or do we need to remain cautious? How do you envision its future?

Generative AI, particularly in well-known models like OpenAI's ChatGPT, Microsoft Copilot, and Alibaba Cloud's Tongyi Qianwen, has showcased remarkable capabilities in generating human-like text, images, and even code. The potential impact of generative AI on various industries is substantial, with robust applications in process automation & innovative problem-solving. However, we do need to take into consideration and practice caution when implementing generative AI into business applications.

As with any powerful technology, ethical considerations are paramount. Ensuring responsible use and preventing the generation of harmful or misleading content is crucial. There's a risk of malicious use, such as generating deep fakes or deceptive content. Robust security measures and awareness are essential to prevent misuse. Striking a balance between innovation and ethical considerations will be key to harnessing the full potential of this technology for a positive impact across industries.

Q. Where does CHARLES & KEITH leverage AI the most?

Given the evolving nature of AI technology and how it aligns with business strategy, we are exploring the best ways to implement AI across our business functions. A general trend we've observed within the fashion industry is using AI to solve common "low-hanging fruit" business processes. For example, AI-powered chatbots and virtual assistants enhance customer support by providing quick responses to inquiries, assisting with product information, and improving overall customer experience.

Another area where the fashion industry can leverage AI is data analysis. AI-powered data analysis enables companies to derive insights from large datasets, understand market trends, and make data-driven decisions.

Q. What other experiences do you foresee technology will bring to fashion retail in the near future? Would you like to share some of your personal ideas on emerging technologies?

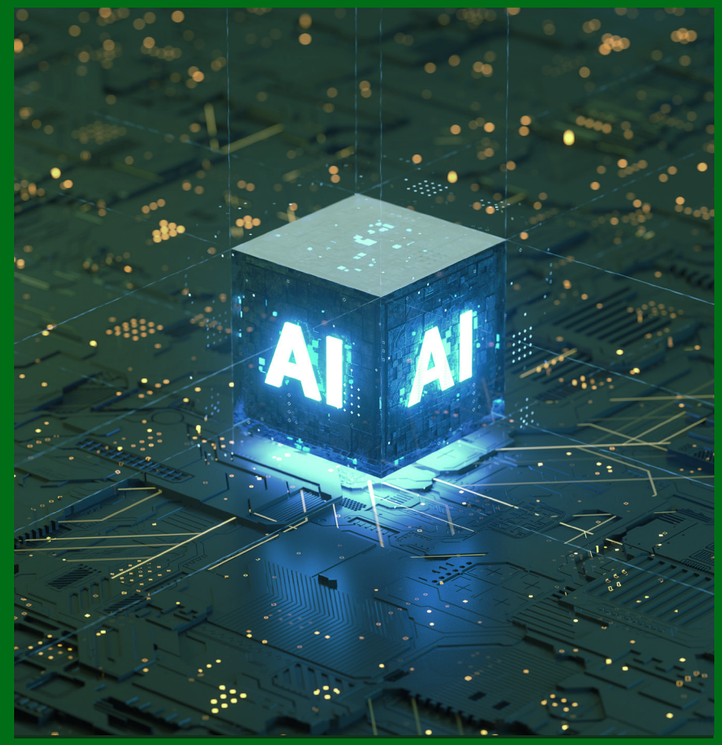
The future holds exciting possibilities for technological advancements that will transform consumer experiences in the fashion retail space. Here are some futuristic experiences that I think will impact the industry:

- AI-driven personalization could reach new heights, with algorithms predicting consumer needs and preferences in real-time. Products and services could be tailored to individual tastes, creating highly personalized and relevant experiences.
- Virtual try-on experiences will become more sophisticated with advanced augmented reality (AR) technology. Customers can virtually try on clothing and accessories in real-time, accurately simulating how items will look on them.
- Smart mirrors equipped with AR interfaces will provide interactive fitting room experiences. Customers can visualize different outfits, access product information, and even receive personalized styling suggestions while trying on clothes.
- Advanced predictive analytics could anticipate consumer needs before they are expressed. Retailers might proactively suggest products, services, or even experiences based on historical data and behavioral patterns.

07 IN CONVERSATION WITH YANG YOU: TRENDS AND CHALLENGES IN THE INTELLIGENT INDUSTRY WITH EFFICIENT AI

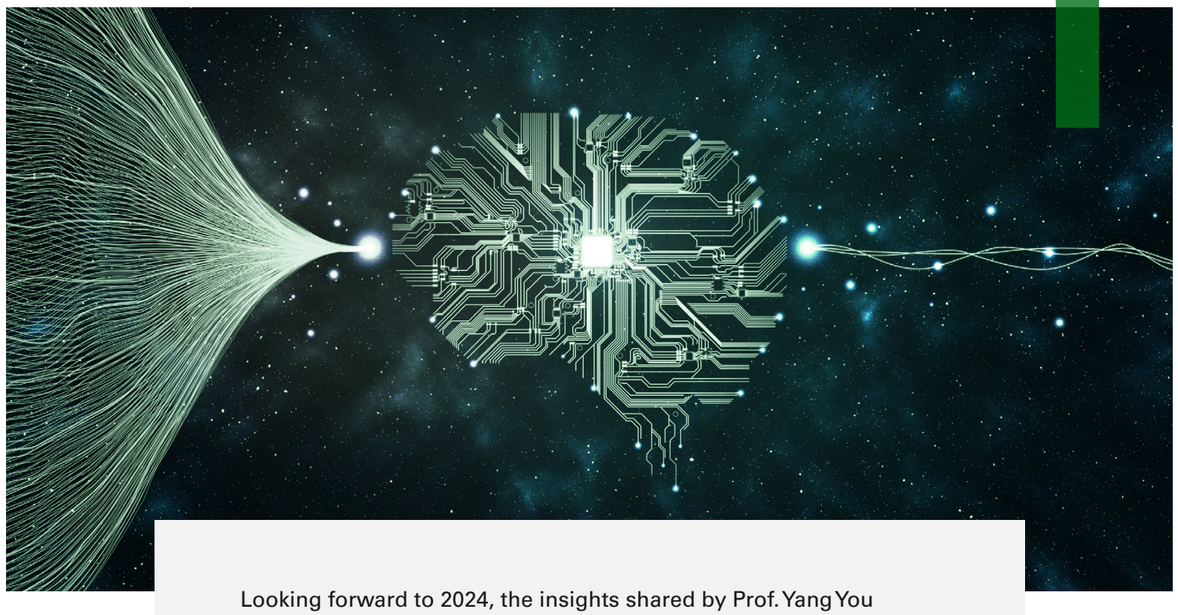
Yang You is a Presidential Young Professor at NUS and the Founder of HPC-AI Tech. He received his Ph.D. in computer science from UC Berkeley. His current research focuses on scaling up deep neural networks training on distributed systems or supercomputers. He was nominated by UC Berkeley for the ACM Doctoral Dissertation Award. He also made Forbes' 30 Under 30 Asia list (2021) and IEEE-CS TCHPC Early Career Award.

Yang You
Presidential Young Professor at NUS and the Founder of HPC-AI Tech



As we conclude the retrospective on the transformative year that was 2023 in the realm of AI large models, it becomes evident that the field has not only thrived but has set the stage for even more remarkable developments in 2024.

Looking back on 2023, the field of AI large models exhibited a flourishing state, with the development speed of these models surpassing imagination. This year, hundreds of enterprises launched their own general-purpose or industry-specific large foundation models. Various industries eagerly integrated these models into their operations and production, utilizing their high efficiency to enhance productivity. From the widespread adoption of ChatGPT to the groundbreaking release of the LLaMA model, the year showcased an array of breakthroughs, pushing the boundaries of what was previously thought possible in the field of AI.



Looking forward to 2024, the insights shared by Prof. Yang You offer a glimpse into the trajectory of AI's evolution. The emphasis on increased standardization and reduced training costs sets the tone for a more accessible and streamlined AI landscape. The critical components of computing power, algorithms, and data are spotlighted as essential pillars for the continued advancement of large models.

Computing power, serving as the foundation for both the training and inference of large models, has gained significant attention during the development of AI. Despite continuous breakthroughs in GPU performance, computing power remains relatively scarce when compared to the computational resources required by models, making it a significant factor contributing to the high costs of large models. Therefore, an efficient AI infrastructure system is essential in the era of large models.

The Colossal-AI team led by Prof. Yang You focuses on precisely combining HPC with AI to complete the training and inference of large models efficiently and cost-effectively. Utilizing parallel computing and heterogeneous memory in HPC, the team can speed up training and reasoning several times, reducing the cost of large models significantly.

In the future, AI infrastructure may face challenges related to the diversification of computing power, such as mixing different chips during training or adapting to various inference chips in onboard scenarios. This will require closer cooperation among computing power vendors, large modeling companies, and open-source community developers to usher in a new wave of large models.

At the algorithmic level, Prof. Yang You expects that in 2024, large models will no longer blindly pursue extremely massive scales, while models from the open-source community, such as the LLaMA architecture, will be more widely used. With the assistance of MoE (Mixture of Experts) technology, achieving similar effects to super-large models without their scale will be feasible across a wide range of industry demands like finance, office, and education. Training such models can significantly save costs and lower the subsequent fine-tuning and iteration costs. In 2024, there will be ample market development space for these models and more industry-specific models will be implemented.

Industrial data is indispensable to the implementation of industrial models. However, industrial data is often confidential and is unique to each business. Consequently, more companies will opt to train large models themselves to ensure data confidentiality. For non-Internet-based enterprises, starting from scratch is challenging. If there is a complete training solution coupled with private computing power, companies can focus solely on handling the data, facilitating the rapid customization of private large models. Prof. Yang You envisions a future where all-in-one private training solutions simplify the large model training process, making it more accessible and enabling companies to focus on data handling and rapid customization.

In conclusion, the journey from 2023 to 2024 in the AI large models landscape is marked by a commitment to accessibility, efficiency, and collaboration. As a technological innovation hub in Southeast Asia, Singapore's demand and support for AI are also growing. Colossal-AI, as a partner of Alibaba Cloud Singapore, looks forward to discussing the development and opportunities of enterprise-level AI applications in the Singapore market under the wave of generative AI with Alibaba Cloud.

NEXUS

Through our NEXUS CLOUD, we are committed to openness and inclusivity in the AI era through our collaborative endeavors. “Nexus” represents partnerships, business collaborations, learning initiatives, and social integration, envisioning a future where AI empowers ecosystems and connects society.

By engaging a global network of tens of thousands of partners, fostering a vibrant digital ecosystem, and prioritizing community engagement and talent empowerment, Alibaba Cloud envisions a connected world where AI bridges gaps and drives collective growth.

08 UNLEASHING THE POWER OF GPT/LLMS FOR DATA CENTER EVOLUTION

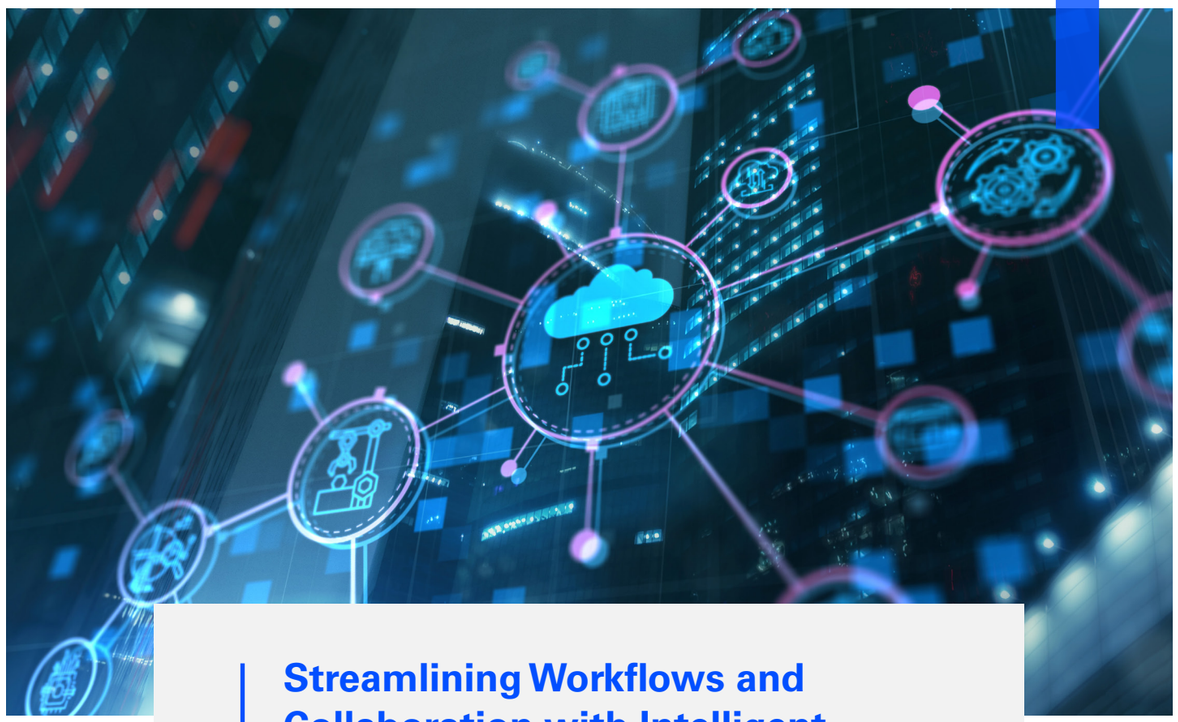
The imperative for digital transformation has taken center stage in the ever-evolving landscape of data centers. Traditional approaches encounter constraints, prompting the industry to adopt innovative solutions. Enter the Generative Pre-trained Transformer/Large Language Models (GPT/LLMs), a groundbreaking solution to surmount the limitations of traditional approaches. This cutting-edge technology harnesses the power of intelligent automation and predictive analytics, revolutionizing how data centers undergo digital transformation.

Yonggang Wen
Professor & President's Chair,
School of Computer Science and Engineering at Nanyang Technological University (NTU), Singapore

Sam Ho
Head of Product,
Red Dot Analytics

For the data center industry, GPT/LLMs have garnered attention for their potential to reduce human risk, predict power outages, minimize hotspots, and optimize cooling practices. GPT/LLMs' transformative impact on data center digital transformation would stem from four key capabilities:

- Intelligent Automation
- Empowering Decision-Making
- Predictive Analytics
- Digital Twin Development



Streamlining Workflows and Collaboration with Intelligent Automation

My colleagues and I at Red Dot Analytics envision a future where the digital transformation of data center operations is further accelerated through the powerful capabilities of GPT/LLMs. Building upon our achievements, which have already reduced human risk by 70%, Red Dot Analytics aims to mark even more significant milestones in the industry.

GPT/LLMs' intelligent automation brings benefits such as streamlined workflow automation. Manual processes, such as the ticketing system, can be handled seamlessly by the GPT/LLMs - categorizing tickets, assigning priorities, and suggesting solutions based on data and best practices. This saves time, minimizes errors, and ensures swift issue resolution.

Collaboration within data center facility management can be enhanced through GPT/LLMs that have been fine-tuned over domain knowledge and operational datasets. GPT/LLMs facilitate seamless communication and quick access to accurate information for teams. They further assist maintenance teams in enhancing communication by transforming intricate technical information into easily comprehensible language. This feature proves valuable for organizations with geographically diverse teams, guaranteeing equal access to information and promoting effective collaboration on maintenance tasks.

Unlocking Efficiency and Sustainability through Empowered Decision-Making

By tapping into its vast knowledge repository, GPT/LLMs provide valuable insights for informed decision-making, particularly regarding energy optimization. GPT/LLMs can analyze historical and real-time data to identify energy-saving opportunities, recommending adjustments such as chiller set points, airflow management, and server rack arrangement. These optimizations can lead to significant energy savings, reduced costs, and minimized environmental impact.

GPT/LLMs can analyze factors such as ambient temperature, equipment workload, and cooling system efficiency to suggest adjustments for optimal cooling practices. This ensures that data center operators can strike the right balance between maintaining optimal operating temperatures for equipment and minimizing energy consumption, improving overall efficiency, and extending the lifespan of critical infrastructure.

Red Dot Analytics aims to leverage GPT/LLMs' capabilities in analyzing energy consumption and optimizing cooling practices and develop a plug-in application to achieve even more excellent results in helping data centers, with a vision to surpass the current achievement of up to 40% energy reduction.

Strategic Insights for Proactive Maintenance through Predictive Analytic

Predicting and proactively addressing potential issues is crucial for data center operators. Traditional analytics fell short in extracting insights, but GPT/LLMs' predictive analytics capabilities introduced a new era of prescriptive maintenance. GPT/LLMs' predictive analytics can optimize resource allocation by analyzing equipment performance, workloads, and environmental conditions.

Operators can make informed decisions about capacity planning, load balancing, and upgrades, improving efficiency and reducing costs. GPT/LLMs can continuously monitor data center metrics, identifying potential hotspots, equipment malfunctions, or issues for early anomaly detection. GPT/LLMs' ability to process and analyze unstructured data is a major advantage in predictive maintenance. GPT/LLMs' language processing capabilities can extract valuable insights from sources such as maintenance logs & technician notes, enabling a comprehensive understanding of equipment health and performance.

Instead of relying on routine, time-based maintenance schedules, GPT/LLMs can prescribe a maintenance plan based on actual equipment conditions and predicted failure risks. This shift from reactive to proactive maintenance ensures that resources are utilized efficiently and that maintenance activities are focused precisely where needed, optimizing operational efficiency and extending the lifespan of critical infrastructure.

Accelerating Digital Twin Development

Data center operators no longer need to be hindered by high costs, extensive maintenance, and lengthy development times for digital twin implementation. GPT/LLMs emerge as the transformative solution, accelerating digital twin development and unlocking its full potential within the data center. With GPT/LLMs, data center operators can rapidly generate highly accurate digital twins by leveraging architectural designs, equipment specifications, and operational data.

Businesses no longer need to rely on manual modeling techniques and complex software for digital twin development. Red Dot Analytics is at the forefront of developing a groundbreaking solution for building digital twins. With an innovative approach, we can translate instructions and prompts into generating layouts, equipment arrangements, and simulations at desired setpoints, transforming traditional methodologies. We capture the intricate details of the physical infrastructure and its dynamic interactions, empowering operators to engage in virtual testing, optimization, and predictive modeling.

GPT/LLMs empower operators to engage in virtual testing, optimization, and predictive modeling within the digital twin environment. It processes vast amounts of data, learns from historical patterns, and enhances the accuracy and reliability of digital twins. This enables operators to design more efficient infrastructures, improve planning, and enhance operational understanding with unprecedented speed and accuracy.

The accelerated development of digital twins enables operators to optimize resource utilization, test energy management strategies, assess carbon emissions, and drive sustainability initiatives. GPT/LLMs also support compliance by providing insights into regulatory requirements and industry standards.