Alibaba Cloud Generative AI Solution Whitepaper

Alibaba Cloud

alibabacloud.com

LEGAL NOTICES

Alibaba Cloud reminds you to carefully read through and completely understand all of the content in this section before you read or use this document. If you read or use this document, it is considered that you have identified and accepted all contents declared in this section.

1. You shall download this document from the official website of Alibaba Cloud or other channels authorized by Alibaba Cloud. This document is only intended for legal and compliant business activities. The contents in this document are confidential, so you shall have the liability of confidentiality. You shall not use or disclose all or part of the contents of this document to any third party without written permission from Alibaba Cloud.
2. Any sector, company, or individual shall not extract, translate, reproduce, spread, or publicize, in any method or any channel, all or part of the contents in this document without written permission from Alibaba Cloud
3. This document may be subject to change without notice due to product upgrades, adjustment, and other reasons. Alibaba Cloud reserves the right to modify the contents in this document without notice and to publish the document in an authorized channel as and when required. You shall focus on the version changes of this document, downloading and acquiring the updated version from channels authorized by Alibaba Cloud.
4. This document is only intended for product and service reference. Alibaba Cloud provides this document for current products and services with current functions, which may be subject to change. Alibaba Cloud provides the document in the context that Alibaba Cloud products and services are provided on an "as is", "with all faults" and "as available" basis. Alibaba Cloud makes the best efforts to provide an

appropriate introduction and operation guide on the basis of current technology, but Alibaba Cloud does not explicitly or implicitly guarantee the accuracy, completeness, suitability, and reliability of this document. Alibaba Cloud does not take any legal liability for any error or loss caused by downloading, using, or putting trust in this document by any sectors, company, or individuals. In any case, Alibaba Cloud does not take any legal liability for any indirect, consequential, punitive, occasional, incidental, or penalized damage, including profit loss due to use of or trust to this document (even if Alibaba Cloud has notified you it is possible to cause this kind of damage).

The responsibilities and liabilities of Alibaba Cloud to its customers are controlled by Alibaba Cloud agreements, and this document is not part of, nor does it modify, any agreement between Alibaba Cloud and its customers.

5. All content including, but not limited to, images, architecture design, page layout, description text, and its intellectual property (including, but not limited to, trademarks, patents, copyrights, and business secrets) used in this document are owned by Alibaba Cloud and/or its affiliates. You shall not use, modify, copy, publicize, change, spread, release, or publish the content from the official website, products, or programs of Alibaba Cloud without the written permission from Alibaba Cloud and/or its affiliates. Nobody shall use, publicize, or reproduce the name of Alibaba Cloud for any marketing, advertisement, promotion, or other purpose (including, but not limited to, a separate or combined form to use the name, brand, logo, pattern, title, product or service name, domain name, illustrated label, symbol, sign, or similar description that may mislead readers and let them identify that it comes from Alibaba Cloud and/or its affiliates, of or from Alibaba Cloud, Aliyun, Wanwang, and/or its affiliates) without the written permission from Alibaba Cloud.
6. If you discover any errors or mistakes within this document, please contact Alibaba Cloud directly to raise this issue.

VERSION HISTORY

January 2025
First Edition – Version 1.0

CONTENTS

1. Trends in Generative AI Solutions 1

1.1 The Beginning of Generative AI Technology 2
1.2 Alibaba Cloud's Role in Shaping Generative AI 4
1.3 Business Challenges and Opportunities 6

2. Alibaba Cloud AI Product Solutions 10

2.1 GenAI Infrastructure - Alibaba Cloud Solutions 12
2.2 GenAI Model - Tongyi Qwen 18
2.3 GenAI Tooling - MaaS Platform with LLMOps 24
2.4 GenAI Apps - Chatbot 31
2.5 GenAI Apps - Document Intelligence 37
2.6 GenAI Apps - BI Analysis 38
2.7 GenAI Apps - Speech 43
2.8 GenAI Apps - Vision 48
2.9 GenAI Apps - Digital Human 53
2.10 GenAI Apps - Code Assistant 55

3. Security and Privacy Are Priorities 60

4. Customer Stories 63

- AstraZeneca with OpenTrek Platform 63
- Haleon with Tongyi Qwen 65
- Shiseido with Tongyi Qwen 68
- X-Verse Technologies with Tongyi Qwen 71
- Lightblue with Tongyi Qwen 75

5. References 78

1. TRENDS IN GENERATIVE AI SOLUTIONS

Generative AI is rapidly transforming industries, with projections indicating it could contribute up to $4.4 trillion annually to the global economy.[1]

The swift evolution of this technology is evident, with significant advancements occurring multiple times a month.

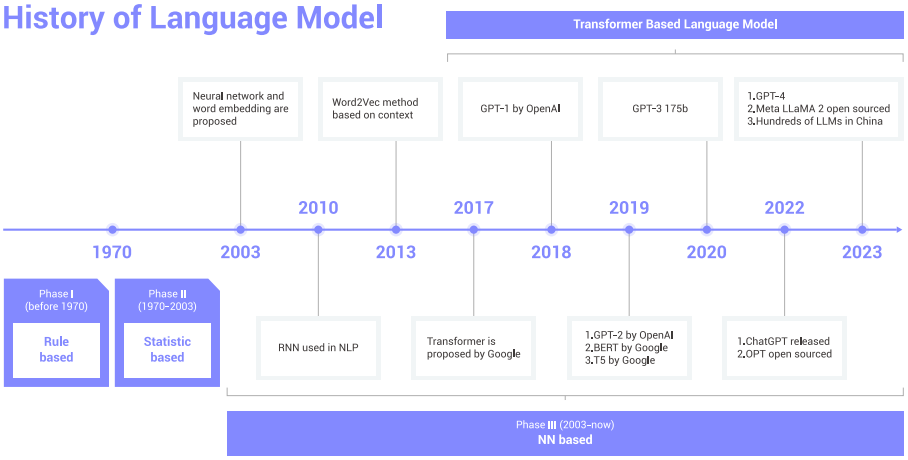
Industries such as banking, high tech, and life sciences are poised to benefit significantly from generative AI, with potential impacts on various job roles and activities.

As generative AI continues to evolve, it is crucial for businesses and policymakers to stay informed and adapt to the changing landscape to harness its full potential while mitigating associated risks.

Generative AI has emerged as one of the most transformative technologies of the modern era, driving innovations across industries with its ability to generate human-like content, automate complex workflows, and enable real-time decision-making. At its core, Generative AI relies on advanced machine learning techniques, such as deep learning and neural networks, to generate content, predict outcomes, and offer intelligent insights. This document explores the evolution of generative AI technologies, key trends in the field, and Alibaba Cloud's contributions to shaping the future of Generative AI.

1.1 THE BEGINNING OF GENERATIVE AI TECHNOLOGY

1.1.1 EVOLUTION OF AI MODELS



Rule-based Systems (1970s-1990s): Early AI systems used fixed instruction sets, limiting their adaptability to new inputs.[2]

Machine Learning (2000s): Enabled systems to learn from data, but traditional models struggled with large-scale, unstructured data.[2]

Deep Learning and Transformers (2010s): Introduced deep learning and Transformer architectures, revolutionizing AI. Models like BERT and GPT enabled human-like text generation.[2]

Generative AI Today: Generative AI now encompasses Large Language Models (LLMs), Vision-Language Models (VLMs), and multimodal systems capable of creating text, images, audio, and even videos.

1.1.2 KEY MILESTONES IN GENERATIVE AI

Large Language Models (LLMs): OpenAI's GPT, Google's PaLM, and Alibaba's Qwen are examples of LLMs excelling at text generation, summarization, translation, and reasoning tasks.

MultiModal LLM (MM-LLM): Multimodal Large Models like Qwen-VL and Qwen-Audio represent the latest advancement in AI, capable of processing and generating data across additional modalities such as images, audio, and video.

Vision-Language Models (VLMs): Models like DALL-E, Stable Diffusion, and Alibaba's AI-driven imaging solutions integrate vision and text for creative and analytical applications.

Retrieval-Augmented Generation (RAG): By combining traditional AI with retrieval mechanisms, RAG enhances accuracy and context by fetching relevant data in real time.[3]

Agents and LLMOps: Emerging trends such as agentic RAG and LLMOps are enabling automation and improved manageability for AI systems in production environments.

1.1.3 OPEN SOURCE VS. CLOSED SOURCE

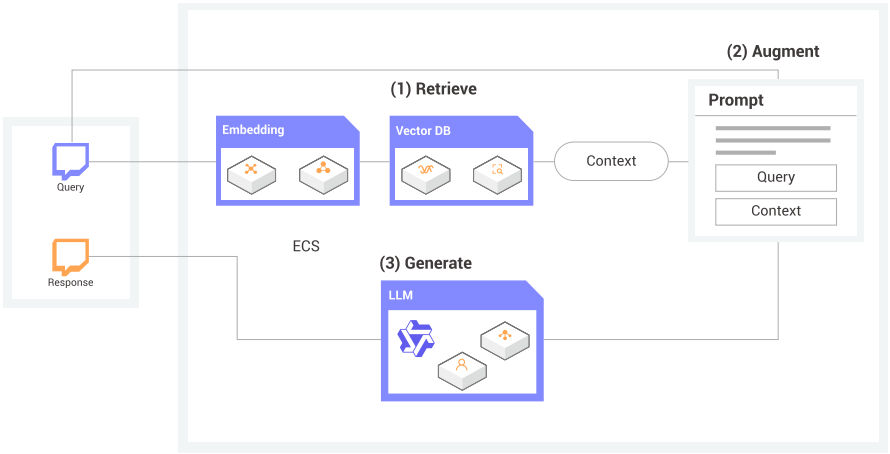
Open Source LLMs: Similar to Qwen family foundation models provide flexibility for developers to fine-tune models. Open-source models foster innovation but require expertise for deployment and optimization.

Closed Source LLMs: Closed Source LLMs like OpenAI's GPT-4 and Alibaba Cloud's Tongyi Qwen offer high performance with integrated solutions and support, ideal for enterprises looking for scalability and security.

1.1.4 RETRIEVAL-AUGMENTED GENERATION (RAG)

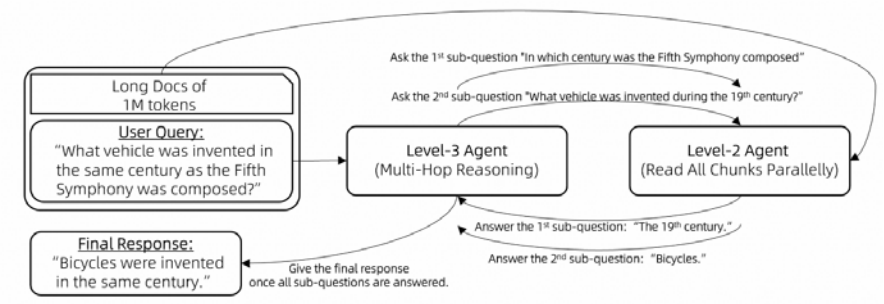
- **The Rise of RAG**

RAG combines the power of LLMs with real-time data retrieval, enabling systems to provide accurate, up-to-date, and contextually relevant responses.



- **Agentic RAG**

Agentic RAG extends traditional RAG by adding agent-like behavior to systems, allowing them to autonomously decide on tasks, fetch data, and generate responses.[4]



Multi-Agent Dataflows of RAG

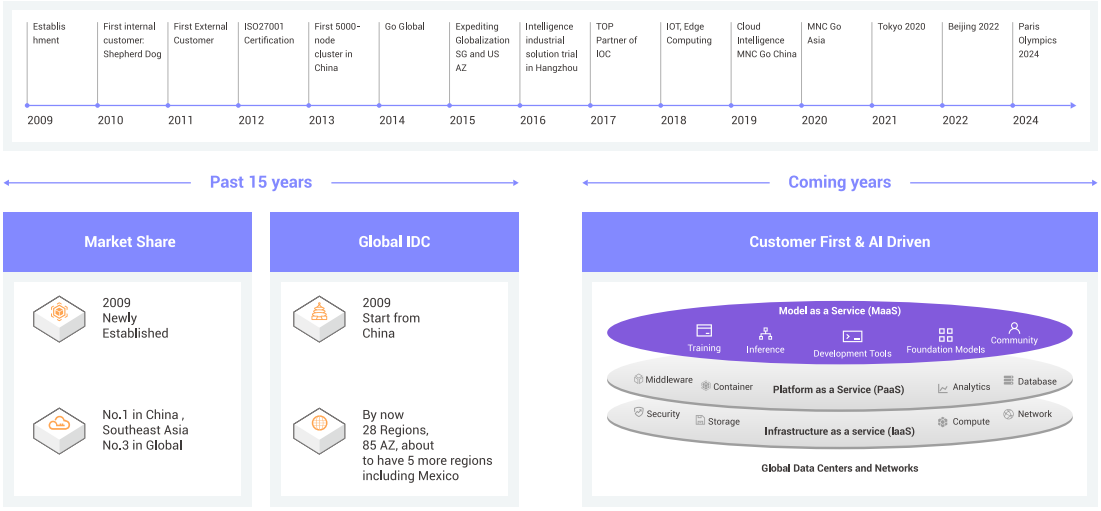
Needle-Bench	(0k, 8k]	(8k, 32k]	(32k, 128k]	(128k, 256k]
32k-Model	87.50	81.19	46.28	0.41
4k-RAG	85.75	78.43	73.26	70.00
4k-Agent	85.41	85.43	85.52	81.82

LV-Eval	(0k, 16k]	(16k, 32k]	(32k, 64k]	(64k, 128k]	(128k, 256k]
32k-Model	49.06	45.72	29.17	12.01	1.32
4k-RAG	49.28	48.90	49.14	49.24	45.80
4k-Agent	51.61	51.52	52.15	51.37	46.60

(Blue = Significantly Better than 32k-Model; Red = Significantly Worse than 32k-Model)

Comparison among the Native Model, RAG and Agentic RAG

1.2 ALIBABA CLOUD'S ROLE IN SHAPING GENERATIVE AI



Alibaba Cloud LLM Solutions

Tongyi Qwen: Alibaba Cloud's flagship LLM excels in understanding and generating human-like text. It supports multiple languages and is optimized for enterprise use cases such as chatbots, document summarization, and creative writing.

Model Studio: A powerful tool for training and deploying LLMs, enabling users to fine-tune models on custom datasets with minimal technical overhead.

Vision-Language Models (VLMs)

Trends: VLMs combine text and image understanding, unlocking applications in e-commerce, media, and healthcare.

Alibaba Cloud Products: Solutions like Wanx provide automated content generation, enabling businesses to create high-quality visuals efficiently.

Alibaba Cloud's RAG Solutions

Vector Databases: Alibaba Cloud's vector database integrates seamlessly with LLMs, ensuring high-speed retrieval of relevant data.

Qwen-Agent: A fully managed agent solution designed to leverage RAG for intelligent workflows and real-time decision-making.

Vector Databases and Embeddings

Vector databases store and retrieve embeddings, enabling efficient similarity searches for AI applications.

Alibaba Cloud Vector Databases: Optimized for speed and scalability, these databases are integral to RAG workflows, recommendation systems, and semantic searches.

AI Agents

AI agents powered by LLMs are becoming central to enterprise automation.

Qwen-Agent: Alibaba Cloud's intelligent agent leverages LLMs and RAG to autonomously manage tasks, improving efficiency across domains like customer support and operations.

LLMOps: Operationalizing Generative AI

LLMOps focuses on the lifecycle management of LLMs, from development to deployment and monitoring.

Alibaba Cloud's Contributions:

Model Studio: Simplifies training and deployment of LLMs.

OpenTrek: Provides tools for collaborative model development and optimization.

AutoML: Automates hyperparameter tuning and feature engineering for large-scale AI projects.

Multimodal AI

Trend: Multimodal AI integrates text, image, audio, and video processing capabilities into unified systems.

Alibaba Cloud Multimodal Solutions: AI-powered products like Qwen-Audio (for audio transcription) and Digital Human (for video content) demonstrate cutting-edge applications in multimodal AI.

Emerging Applications

Generative AI in Code (Lingma): Automating software development with AI-assisted code generation and debugging.

Generative AI in Video (Digital Human): Creating realistic avatars and enabling virtual influencers for the media and entertainment industry.

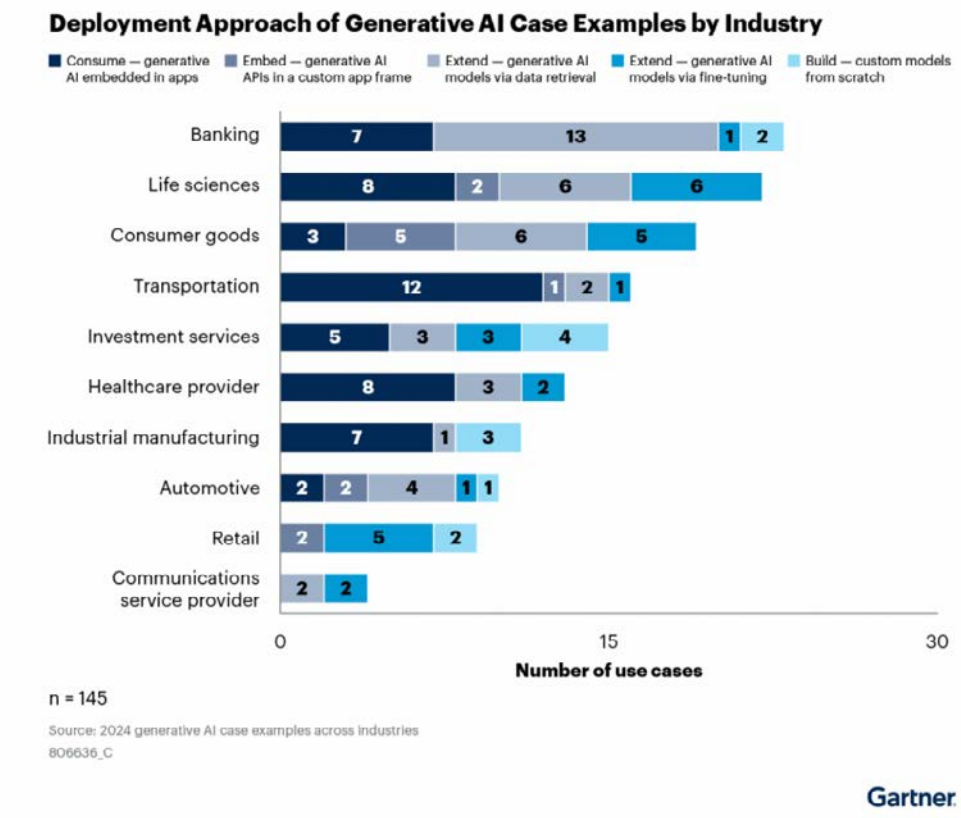
1.3 BUSINESS CHALLENGES AND OPPORTUNITIES

Current Challenges Faced by Enterprises

1. Slow Innovation and Product Development : In a highly competitive market, businesses need to continuously introduce new products and services to meet customer demands. However, traditional R&D processes are often time-consuming and costly, leading to slow progress in innovation and product development.

2. Data Scarcity and Bias Issues : Many industries face challenges with insufficient or poor-quality data, especially in sectors like healthcare and finance. Additionally, existing datasets may contain biases that affect decision-making accuracy and fairness.
3. Inefficient Content Creation and Management : Businesses invest significant resources in content creation, editing, and management, but traditional methods are often inefficient and struggle to produce high-quality content at scale.
4. Insufficient Customer Experience and Personalized Services : As consumers increasingly demand personalized services, companies need to deliver more precise and tailored marketing messages and customer service. However, current technologies often fall short in achieving this goal effectively.
5. Talent Shortage and Skill Enhancement Difficulties : During digital transformation, enterprises often face talent shortages and lack the technical skills needed, particularly in fostering employee collaboration with AI to boost overall productivity.
6. Risk Management and Sustainability Pressures : Companies must identify and manage potential risks in complex market environments while also addressing stringent environmental and sustainability requirements.

Scenarios for Alibaba Cloud Generative AI to Generate Business Value

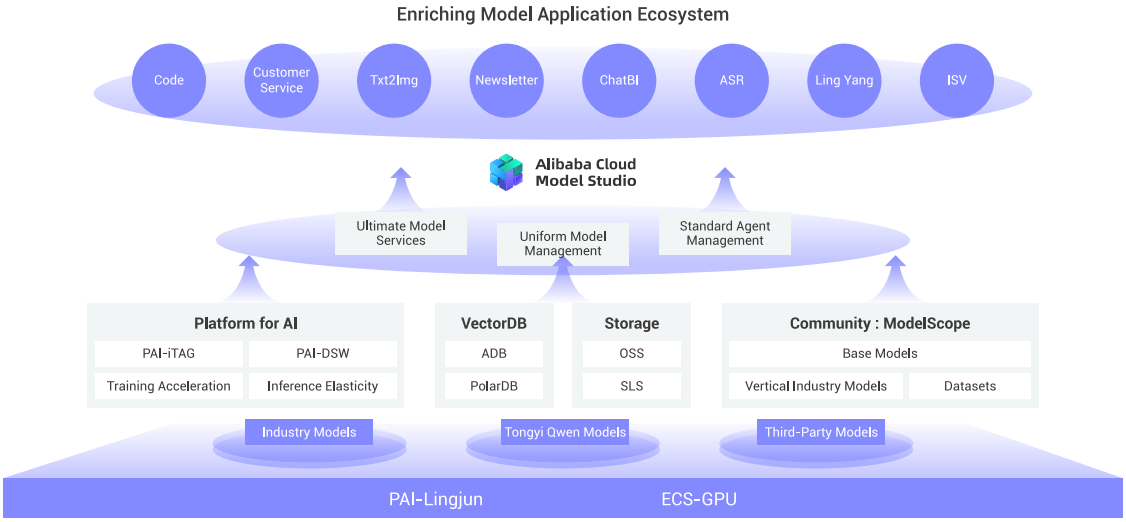


Business Scenarios	Description	Alibaba GenAI Solutions
Customer Service	A retail company needs to handle thousands of customer inquiries daily through its website and mobile app, requiring fast and accurate responses to maintain customer satisfaction. The company aims to reduce response times and improve the quality of interactions without increasing operational costs.	GenAI - Chatbot
Form Automation	A financial institution processes vast amounts of documents daily, including loan applications, contracts, and compliance reports. The manual review process is time-consuming and error-prone, leading to delays and potential regulatory issues. The institution seeks an automated solution to streamline document processing and ensure compliance.	GenAI - Document Intelligence
BI Copilot	An e-commerce platform wants to gain deeper insights into customer behavior and preferences using historical sales data, social media posts, and online reviews. The company aims to generate actionable insights to optimize marketing strategies and improve customer engagement. However, analyzing such diverse and large datasets manually is impractical.	GenAI - BI Analysis
Brand Marketing	A luxury brand plans to launch a virtual showroom where customers can interact with a digital representative who can provide personalized recommendations and answer questions about products. The brand wants to create an immersive and engaging experience that reflects its premium image.	GenAI - Digital Human GenAI - Speech
Software Quality Management	A software development firm struggles with maintaining code quality and consistency across multiple projects due to rapid growth and frequent staff turnover. The firm requires a tool that can assist developers in generating, reviewing, and optimizing code efficiently to reduce errors and enhance productivity.	GenAI - Code Assistant

Business Scenarios	Description	Alibaba GenAI Solutions
Multimodal Content Generation	Travel agencies develop apps offering spoken descriptions and visual cues of tourist attractions which generated by AI, aiming to enhance customer engagement and satisfaction through an interactive experience.	GenAI - Vision GenAI - Speech
Customer Operations	A marketing agency needs to create compelling ad copy, social media posts, and email campaigns for various clients, each with unique branding guidelines and target audiences. The agency aims to produce high-quality content quickly and cost-effectively while ensuring consistency and relevance.	GenAI - Vision
Product R&D	A manufacturing company aims to improve its design processes by leveraging generative AI to explore innovative solutions and optimize designs based on specific goals and constraints. The company wants to reduce the time spent on design iterations and increase the number of viable options available for engineers to evaluate.	GenAI - Code Assistant
Personalized learning and training	A telecom provider faces challenges in training and upskilling its workforce to keep pace with technological advancements and changing customer expectations. The company seeks a solution that can offer personalized learning experiences and support employees in acquiring new skills efficiently.	GenAI - Chatbot

2. ALIBABA CLOUD AI PRODUCT SOLUTIONS

Model as a Service (MaaS): A Paradigm Shift for Model-Centric AI



Alibaba Cloud provides comprehensive support for the training, deployment, and inference of AI models. From the underlying infrastructure to the model layer and up to the application layer, Alibaba Cloud offers reliable, high-performance services. This enables businesses to efficiently achieve innovation in AI applications.

GenAI Infrastructure Layer

- **High-Performance Computing Resources :** Offers cost-effective GPUs and specialized hardware optimized for deep learning tasks.
- **Elastic Scalability :** Dynamically adjust computing resources based on actual needs, avoiding significant upfront hardware investments.
- **Pay-as-You-Go Pricing :** No need to purchase and maintain hardware; pay only for what you use, significantly reducing costs.

Alibaba Cloud provides:

- **Elastic GPU Service :** An IaaS solution providing scalable GPU resources for deep learning, video processing, and scientific computing, ideal for users needing direct hardware control.
- **PAI-Lingjun :** A PaaS platform offering high-performance AI computing with Serverless and Exclusive options, perfect for users seeking advanced AI capabilities without managing infrastructure.

GenAI Model Layer

- Provides official API interfaces for Qwen commercial edition.
- Supports mainstream third-party large models covering various modalities such as text, images, audio, and video.

Alibaba Cloud provides:

- **Tongyi Qwen Models:** A top-tier multi-modal AI model, pre-trained on up to 18 trillion tokens, excelling in text, vision, and audio tasks with enhanced performance in knowledge, coding, and mathematics—serving as the cornerstone for enterprise AI transformation, fully customizable and deployable via Alibaba Cloud Model Studio.

GenAI Tooling Layer

- **Knowledge Base Management :** Facilitates the management and updating of knowledge data within applications.
- **Function Invocation :** Simplifies the integration of complex functionalities, improving development efficiency.
- **Workflow Orchestration :** Supports multi-step workflow design, ensuring clear application logic.
- **Model Customization :** Allows customization of models based on specific business needs, enhancing application performance.

Alibaba Cloud provides:

- **Model Studio :** A streamlined platform for rapid AI application development, offering customizable foundation models like Qwen-Max and advanced features such as Q&A and NL2SQL, ideal for businesses needing quick deployment.
- **OpenTrek :** An enterprise-grade platform for comprehensive AI services, providing private knowledge management, model training, and multi-tenant resource management, tailored for large-scale enterprises requiring extensive AI functionalities and flexible workflows.

GenAI Application Layer

- Enables diverse AI-driven applications across multiple industries, including customer service, document analysis, business intelligence, speech recognition, visual inspection, virtual assistants, and code generation .

Alibaba Cloud provides:

- **Chatbot :** AI-powered conversational agents for customer service and engagement.
- **Document Intelligence :** Automated document processing and analysis for improved efficiency.
- **BI Analysis :** Business intelligence tools for data-driven decision-making.
- **Speech :** Advanced speech recognition and synthesis for voice-based applications.
- **Vision :** Image and video analysis for visual understanding and synthetic content creation using generative AI techniques.
- **Digital Human :** Virtual avatars for interactive and immersive user experiences.
- **Code Assistant :** AI-driven code generation and debugging tools for developers.

2.1 GENAI INFRASTRUCTURE - ALIBABA CLOUD SOLUTIONS

Alibaba Cloud's tailored GenAI-GPU solutions are designed to provide enterprises with the powerful and flexible computing resources necessary for advanced artificial intelligence tasks. These solutions integrate seamlessly with Alibaba Cloud's comprehensive suite of cloud services, enabling businesses to deploy, manage, and scale AI applications efficiently.

2.1.1 ELASTIC GPU SERVICE - ELEVATE YOUR AI CAPABILITIES WITH ALIBABA CLOUD

As industries look to enhance their data processing and AI capabilities, Alibaba Cloud's Elastic GPU Service emerges as a crucial resource. Designed for businesses driving innovation through AI, high-performance computing, and graphics-intensive applications, this service combines the power of GPUs with the flexibility and scalability of cloud computing.

Elastic GPU Service Overview

Elastic GPU Service provides a complete service system that combines software and hardware to help you flexibly allocate resources, elastically scale your system, improve computing power, and lower the cost of your AI-related business. It applies to scenarios (such as deep learning, video encoding and decoding, video processing, scientific computing, graphical visualization, and cloud gaming).

Elastic GPU Service provides GPU-accelerated computing capabilities and ready-to-use, scalable GPU computing resources. GPUs have unique advantages in performing mathematical and geometric computing, especially floating-point and parallel computing. GPUs provide 100 times the computing power of their CPU counterparts.

Key Features of Elastic GPU Service

The following table compares GPU-accelerated instances that are provided by Elastic GPU Service and self-managed GPU-accelerated servers.

Item	GPU-accelerated instance	Self-managed GPU-accelerated server
Flexibility	• Allows you to create one or more GPU-accelerated instances with ease. • Supports flexible changes between instance specifications that are configured with different vCPUs, GPUs, and memory, including online upgrades and downgrades. • Provides adjustable bandwidths.	• Requires an extended subscription period. • Provides server specifications that cannot be changed. • Requires a one-off purchase of a bandwidth that cannot be adjusted.
Ease of use	• Provides online web management tools that are easy and convenient to use. • Provides built-in mainstream operating systems, such as an activated genuine Windows operating system, and supports online switching between operating systems. • Allows you to purchase and install GPU drivers when you purchase the instance.	• Does not provide online management tools and requires complex maintenance. • Requires you to install and replace an operating system on your own. • Requires you to purchase and install GPU drivers on your own.

Item	GPU-accelerated instance	Self-managed GPU-accelerated server
Disaster recovery and backup	• Uses a triplicate storage mechanism by which three copies of each piece of data are stored. When one copy is corrupted, the data can be restored from the another copy within a short period of time. • Allows hardware to be automatically recovered when failures occur.	• Requires you to build your disaster recovery environment on your own by using high-cost conventional storage devices. • Requires you to manually restore corrupted data.
Security	• Effectively defends against Media Access Control (MAC) spoofing and Address Resolution Protocol (ARP) attacks. • Defends against DDoS attacks by using blackhole filtering and traffic scrubbing. • Provides additional services such as scanning for port intrusions, trojans, and vulnerabilities.	• Poorly defends against MAC spoofing and ARP attacks. • Requires traffic scrubbing and blackhole filtering devices at additional costs. • Usually encounters problems such as port scans, trojans, and vulnerabilities.
Cost	• Supports the subscription and pay-as-you-go billing methods. You can select an appropriate billing method based on your business requirements. • Allows you to purchase on-demand resources without the need to make a large upfront investment.	• Requires you to purchase resources by paying upfront to meet configuration requirements during peak hours. • Requires a large upfront investment and results in serious resource waste.

Varied Computing Capabilities

Elastic GPU Service has a large number of arithmetic logic units (ALUs) that can be used for large-scale parallel computing. It uses the latest GPU acceleration chips and provides various accelerator cards (such as FPGA–Field-Programmable Gate Array, GPU, and ASIC–Application-Specific Integrated Circuit) to serve business purposes (such as AI, graphics, transcoding, and encryption).

Ease of Use

GPU resources are globally deployed across different geographical locations. Simple logic control units allow you to scale your system based on your business requirements. Elastic GPU Service also provides auxiliary tools (like AIACC, FastGPU, and cGPU).

High Network Performance

Elastic GPU Service uses the SHENLONG architecture to improve server performance and reduce I/O latency. GPU supports up to 24 million pps, a bandwidth of up to 64 Gbit/s over VPCs, and an 800G RDMA network. It is suitable for high-throughput scenarios where multiple threads run in parallel to process computing tasks.

Superior Computing and Network Performance

- **High Computing Power:** The service, when combined with high-performance CPUs, delivers mixed-precision processing capabilities at impressive magnitudes, offering up to 1,000 trillion TFLOPS.
- **Optimized Network:** Leverages Virtual Private Clouds (VPCs) for efficient data transmission, supporting high-bandwidth needs up to 50 Gbit/s using the SCC network, ideal for low-latency operations.

Flexible Purchasing Options

Customers can select from various billing methods tailored to their needs, be it subscription-based, pay-as-you-go, or leveraging cost-saving options like preemptible instances and reserved resource plans.

DeepGPU for Enhanced GPU Resource Management

Alibaba Cloud's DeepGPU framework enhances GPU resource management with efficient utilities such as:

- **ACSpeed and AGSpeed:** Offering optimization for distributed AI training and PyTorch model performance.
- **cGPU Technology:** Allowing the fragmentation and allocation of a single GPU across multiple isolated containers.
- **FastGPU Toolkit:** Simplifying the building of AI computing tasks with intuitive interfaces and command-line options.

GPU-Accelerated Instance Families

Alibaba Cloud's ECS (Elastic Compute Service) offers diverse instance families tailored for GPU acceleration. These include the enterprise-level heterogeneous computing, ECS Bare Metal Instances, and Super Computing Cluster (SCC) categories, allowing for seamless integration and deployment of GPU-accelerated capabilities, just like any common ECS instance..

2.1.2 PAI-LINGJUN – EMPOWERING GENERATIVE AI WITH ADVANCED CLOUD SOLUTIONS

PAI-Lingjun is a PaaS service for large-scale deep learning and intelligent computing, available as Serverless on Alibaba Cloud and an Exclusive Edition. It optimizes software and hardware to create a high-performance computing platform for AI, enhancing efficiency in training models, autonomous driving, research, finance, and more.

Serverless

Lingjun Serverless Edition can help you quickly set up and run AI computing tasks. It manages complex heterogeneous systems based on automatic operations and maintenance (O&M), and seamlessly integrates with Alibaba Cloud computing, storage, and network services.

High-Performance RDMA Network

Alibaba Cloud's high-performance Remote Direct Memory Access (RDMA) networks greatly accelerate AI training, with high-speed and low-latency transmission at 800 Gbit/s and GPU direct connection technologies that improve transmission stability and security.

Efficient CPFS Storage System

Cloud Paralleled File System (CPFS) uses a fully parallel storage architecture and supports POSIX/MPI-IO and Network File System (NFS) protocols. A single cluster supports data throughput of up to 2 TB/s and 30 million IOPS, providing efficient and reliable storage services for AI training.

Comprehensive AI Acceleration

Our distributed training acceleration engine provides data set acceleration, computing acceleration, algorithm optimization, scheduling algorithms, and resource optimization. This ensures computing power is fully utilized, comprehensively improving the speed and efficiency of AI training and inference.

Challenges in LLM Development

Developing and deploying LLMs present several challenges:

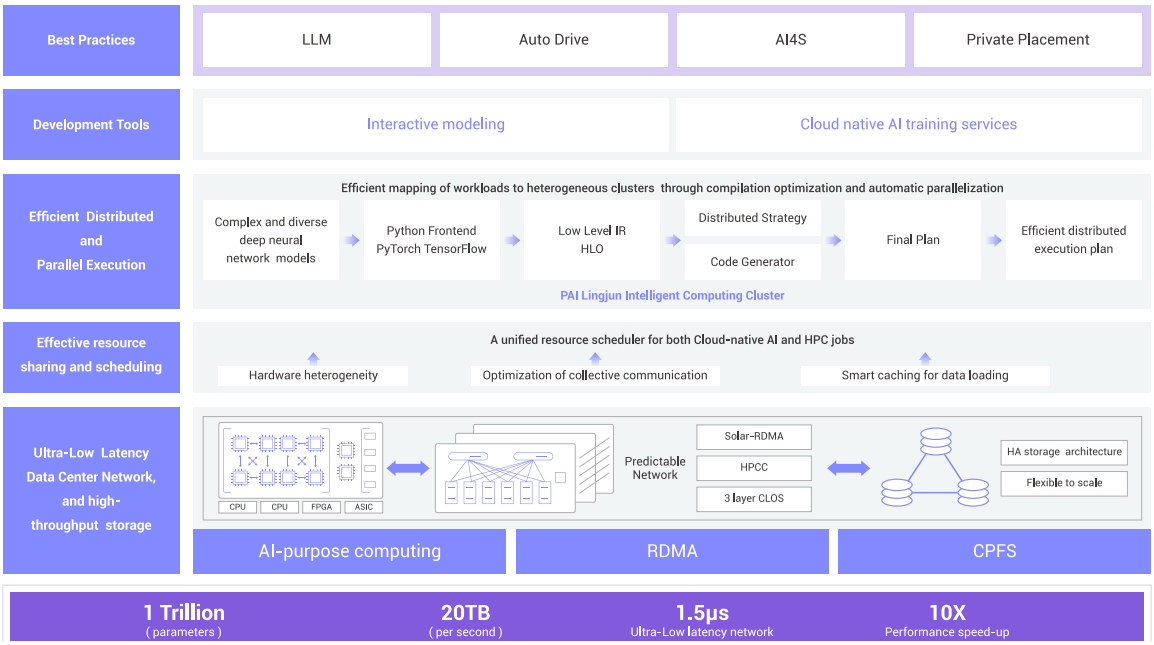
- **Training Tasks Over Multiple Nodes**
  - Ensuring stability, performance, and efficiency across distributed systems is critical. Hardware failures, network issues, and framework bugs can disrupt training processes.

- **Resource Intensive Training**
  - LLMs require vast computational resources, including thousands of GPUs and extended training times spanning months. Optimizing resource utilization is essential for cost-effective training.
- **Inference Optimization**
  - Achieving low-latency online inference and high-throughput offline inference necessitates efficient use of computational resources, often involving hundreds to thousands of GPUs based on query-per-second (QPS) requirements.

PAI-Lingjun's Solution

PAI Lingjun addresses these challenges through a comprehensive suite of tools and services designed to streamline LLM development and deployment:

- **Compute Isolation**
  - Leveraging secure container technology, PAI Lingjun ensures each model instance operates within its own lightweight virtual machine sandbox. This isolation prevents interference between instances, ensuring data storage security and consistent performance.
- **Data Processing Log Auditing**
  - Detailed audit logs track all application interface API calls, VPC operations, and sensitive data access/decryption actions. These logs enable customers to perform in-depth analysis and gain valuable insights into system behavior.
- **Real-Time Content Moderation**
  - PAI Lingjun includes default content safety filters and offers enhanced real-time monitoring for identifying abnormal risks. This feature supports ethical standards and prevents misuse of personal or sensitive data.
- **Model Security Evaluation**
  - For user-fine-tuned models, PAI Lingjun provides specialized security testing, including adversarial testing (Red Team), to assess input/output risk levels. This helps users comply with regulatory requirements and maintain model integrity.

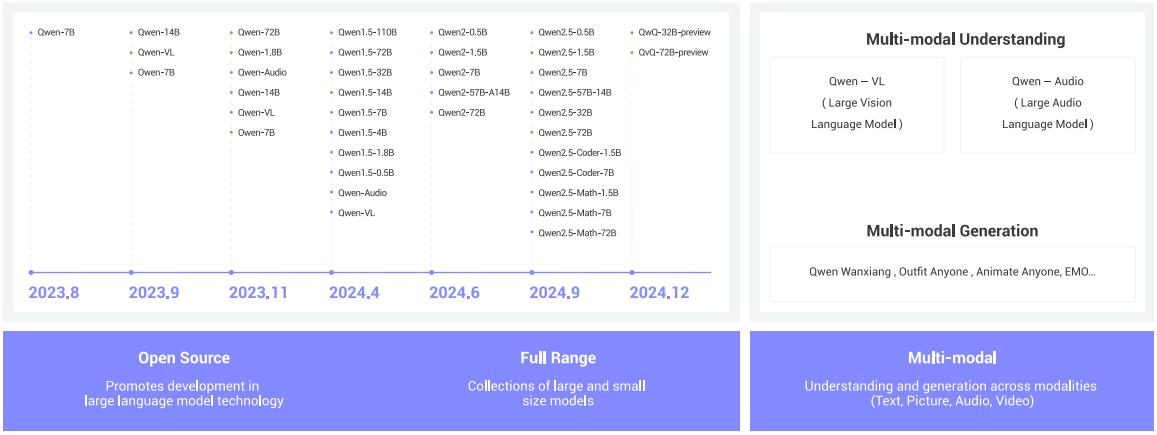


Positioning for Success with Alibaba Cloud

Alibaba Cloud's GenAI-GPU infrastructure enables enterprises to efficiently and securely use high-demand AI capabilities. Products like ECS with GPU and PAI-Lingjun support advancements in technology, finance, healthcare, and more. Partner with us for cutting-edge tech, strategic support, and customer-centric solutions, ensuring readiness for future challenges in the digital landscape.

2.2 GENAI MODEL - TONGYI QWEN

In the rapidly evolving landscape of artificial intelligence, Alibaba Cloud has introduced Tongyi Qwen, an extensive family of large language models (LLMs) and multimodal models that serve as foundational building blocks for generative AI solutions. The Tongyi series is designed to power diverse applications ranging from natural language processing and code generation to vision-language tasks and audio processing.



2.2.1 THE QWEN FOUNDATION MODELS

The Tongyi family encompasses several specialized models tailored for different domains and use cases:

1. Qwen2 Series:

- **Qwen2.5:** A general-purpose LLM with enhanced mathematical reasoning capabilities, making it suitable for applications requiring complex calculations.
- **Qwen2.5-Coder:** Specialized for coding assistance, this model can generate, debug, and optimize code snippets across multiple programming languages.
- **Qwen2-VL:** A multimodal model capable of understanding both text and visual content, opening up possibilities for rich media applications.

2. Qwen Agent:

- An agent framework built on top of Qwen>=2.0, incorporating advanced features such as function calling, code interpretation, retrieval-augmented generation (RAG), and browser extensions to facilitate seamless interaction between users and AI.

3. Qwen Audio & Vision Language Models:

- **Qwen-Audio:** Focused on audio processing, this model can transcribe, translate, and generate speech, enhancing voice-based services and accessibility tools.
- **Qwen-VL:** Designed for vision-language tasks, this model bridges the gap between textual and visual information, supporting applications like image captioning and visual question answering.

4. Supporting Tools and Resources:

- **Qwen Blog:** Provides updates, insights, and discussions about the development and application of the Tongyi models.
- **Qwen Cookbook:** Offers open-source examples and guides for developers looking to build applications using the Tongyi models.

Capabilities and Applications

The Tongyi foundation models offer a wide array of capabilities that cater to various business needs:

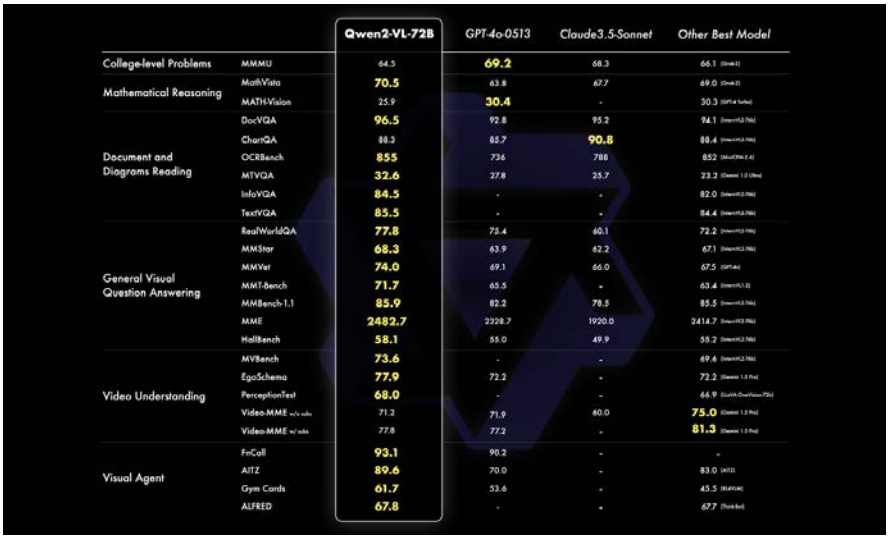
- **Natural Language Understanding and Generation:** With robust NLP skills, Tongyi models can understand context, sentiment, and intent, enabling them to engage in meaningful conversations, answer queries, and create content.
- **Code Generation and Assistance:** Developers can leverage Qwen2.5-Coder to automate coding tasks, enhance productivity, and maintain code quality through intelligent suggestions and corrections.
- **Multimodal Interaction:** By integrating text, image, and audio processing, Tongyi models support rich and immersive user experiences across platforms and devices.
- **Automation and Efficiency:** Automating routine and complex processes reduces human effort and increases operational efficiency without compromising accuracy.
- **Personalization:** Tailoring interactions and content to individual preferences improves user satisfaction and engagement levels.

2.2.2 QWEN2-VL

Qwen2-VL marks a significant milestone in vision-language modeling, offering state-of-the-art performance across various benchmarks.

Key features

- **Advanced Visual Understanding:** Achieves top-tier performance on benchmarks such as MathVista, DocVQA, RealWorldQA, and MTVQA.
- **Long Video Comprehension:** Capable of processing videos over 20 minutes for high-quality video-based question answering, dialog, and content creation.
- **Device Integration:** Functions as an intelligent agent capable of operating mobile phones, robots, and other devices based on visual environment and text instructions.
- **Multilingual Support:** Supports multiple languages inside images, serving a global user base with enhanced multilingual capabilities.



		Qwen2-VL-72B	GPT-4o-0513	Claude3.5-Sonnet	Other Best Model
College-level Problems	MMMU	64.5	**69.2**	68.3	66.1
Mathematical Reasoning	MathVista	**70.5**	63.8	67.7	69.0
	MATH-Vision	25.9	**30.4**	-	30.3
Document and Diagrams Reading	DocVQA	**96.5**	92.8	95.2	94.1
	ChartQA	88.3	85.7	**90.8**	88.4
	OCRBench	**855**	736	788	852
	MTVQA	**32.6**	27.8	25.7	23.2
	InfoVQA	**84.5**	-	-	82.0
	TextVQA	**85.5**	-	-	84.4
General Visual Question Answering	RealWorldQA	**77.8**	75.4	60.1	72.2
	MMStar	**68.3**	63.9	62.2	67.1
	MMVet	**74.0**	69.1	66.0	67.5
	MMT-Bench	**71.7**	65.5	-	63.4
	MMBench-1.1	**85.9**	82.2	78.5	85.5
	MME	**2482.7**	2328.7	1920.0	2414.7
	HallBench	**58.1**	55.0	49.9	55.2
Video Understanding	MVBench	**73.6**	-	-	69.6
	EgoSchema	**77.9**	72.2	-	72.2
	PerceptionTest	**68.0**	-	-	66.9
	Video-MME w/o subs	71.2	71.9	60.0	**75.0**
	Video-MME w/ subs	77.8	77.2	-	**81.3**
Visual Agent	FnCall	**93.1**	90.2	-	-
	ATTZ	**89.6**	70.0	-	83.0
	Gym Cards	**61.7**	53.6	-	45.5
	ALFRED	**67.8**	-	-	67.7

Qwen2-VL showcases top-tier performance across most metrics, often surpassing even closed-source models like GPT-4o and Claude 3.5-Sonnet.

2.2.3 QVQ - QWEN MULTIMODAL REASONING MODEL

Qwen-QVQ is an experimental research model focused on advancing visual reasoning capabilities. It builds upon the foundation of Qwen2-VL but introduces several innovations:

- **Enhanced Multidisciplinary Understanding:** Scores 70.3% on the Multimodal Massive Multi-task Understanding (MMMU) benchmark, demonstrating strong cross-disciplinary reasoning.
- **Improved Mathematical Reasoning:** Shows significant improvements on MathVision benchmarks, particularly in handling complex mathematical problems.
- **Limitations and Challenges:** Acknowledges limitations such as language mixing, recursive reasoning loops, and potential hallucinations during multi-step visual reasoning tasks.

	QVQ 72B-preview	OpenAI o1 2024-12-17	GPT-4o 2024-05-13	Claude3.5 Sonnet	Qwen2-VL 72B
MMMU Val	70.3	77.3	69.1	70.4	64.5
MathVista Test-mini	71.4	71.0	63.8	65.3	70.5
MathVision Full	35.9	-	30.4	35.6	25.9
OlympiadBench Full	20.4	-	25.9	-	11.2

Evaluate Result on 4 datasets [6]

Performance Benchmarks

- **MMMU Benchmark:** Scored 70.3%, showcasing its powerful ability in multidisciplinary understanding and reasoning.
- **MathVista(mini):** Achieved 71.4%, highlighting advancements in mathematical reasoning.
- **MathVision(full):** Demonstrated progress with a score of 35.9%.

Qwen2-VL and Qwen-QVQ represent significant advancements in vision-language modeling, pushing the boundaries of multimodal understanding and reasoning. By integrating these models into various applications, Alibaba Cloud aims to empower businesses and developers to harness the power of generative AI, driving innovation and delivering value across industries. These models not only enhance productivity and decision-making but also pave the way for future developments in AI technology.

2.2.4 QWQ - ADVANCED REASONING MODEL

QwQ (Qwen with Questions) is an advanced AI model designed to explore the depths of thinking, questioning, and understanding. Modeled after an eternal student of wisdom, QwQ approaches every problem—whether in mathematics, coding, or general knowledge—with genuine curiosity and skepticism.

Key Advantages

1. **Curiosity-Driven Learning :** QwQ questions its assumptions and explores multiple paths of thought, leading to a more thorough understanding.
2. **Deep Analytical Capabilities :** Particularly strong in mathematics and programming, QwQ excels in solving complex problems through patient and thoughtful analysis.
3. **Versatile Application :** Suitable for a wide range of applications, from scientific problem-solving to real-world programming scenarios.

	QwQ 32B-preview	OpenAI o1-preview	OpenAI o1-mini	GPT-4o	Claude3.5 Sonnet	Qwen2.5-72B Instruct
GPQA Pass@1	65.2	72.3	60.0	53.6	65.0	49.0
AIME Pass@1	50.0	44.6	56.7	9.3	16.0	23.3
MATH-500 Pass@1	90.6	85.5	90.0	76.6	78.3	82.6
LiveCodeBench 2024.08-2024.11	50.0	53.6	58.0	33.4	36.3	30.4

Benchmark result of QwQ [7]

Core Features

1. **Philosophical Approach :** QwQ's design emphasizes introspection and self-questioning, ensuring that it does not settle on easy answers but seeks deeper insights.
2. **Multimodal Reasoning :** Capable of handling various types of data, including text, code, and mathematical problems, making it versatile across different domains.
3. **Benchmark Performance :** Demonstrates exceptional performance in benchmarks such as GPQA (65.2%), AIME (50.0%), MATH-500 (90.6%), and LiveCodeBench (50.0%).

QwQ represents a significant step forward in AI reasoning capabilities. Its unique approach to learning and problem-solving makes it a powerful tool for tackling complex challenges in mathematics, programming, and beyond. While still in its early stages, QwQ's potential for growth and improvement promises exciting developments in the field of AI.

2.2.5 APPLICATIONS AND USE CASES

Industry	Application Scenario
Education	Visual Question Answering (VQA):Accurately answers questions about images and videos, making it ideal for educational tools and interactive media.
Legal and Financial Services	Document Analysis:Extracts meaningful insights from unstructured documents, including tables and diagrams, aiding industries like legal and financial services.
Mobile and Robotics	Automated Assistance:Integrates with mobile devices and robots to provide real-time assistance based on visual input and text instructions.
Multimedia and Entertainment	Content Creation:Generates high-quality content for multimedia platforms, enhancing user engagement and creativity.

2.2.6 COMMUNITY AND ECOSYSTEM

Alibaba Cloud fosters an active community around the Tongyi models on Gtihub and HuggingFace, encouraging collaboration, innovation, and knowledge sharing.

https://huggingface.co/Qwen

https://github.com/QwenLM

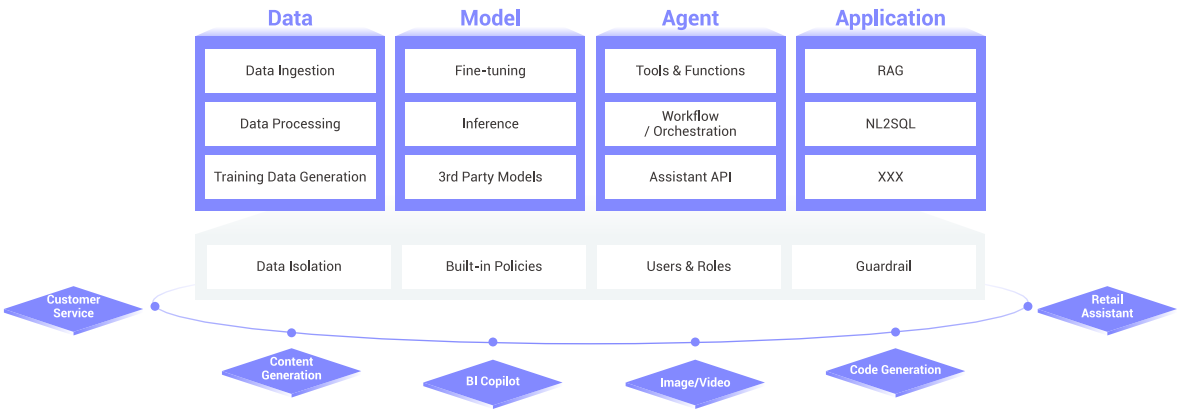
As of November 2024, the Qwen series has seen over **40 million downloads** and **80,000 derivative models**, making it the foundational model with the most derivative models worldwide.

2.3 GENAI TOOLING - MAAS PLATFORM WITH LLMOPS

2.3.1 ALIBABA MODEL STUDIO

Key Features of Model Studio

Model Studio: One-stop LLM Application Development Platform



Why Choose Alibaba Cloud Model Studio?

Alibaba Cloud Model Studio offers an all-in-one platform designed to streamline the development of generative AI applications. The platform provides:

- **Industry-Leading Foundation Models:** Access to state-of-the-art models like Qwen-Max, Qwen-Plus, and Qwen-Turbo, allowing customization with enterprise data through a one-click setup of Retrieval-Augmented Generation (RAG) architecture.
- **Capability-Enhanced Models:** Enhance applications with advanced AI features including Q&A, writing, and NL2SQL functionalities, powered by Alibaba Cloud's proprietary and open-source models.
- **Simplified Development:** Accelerate application development with pre-built workflows, visual orchestration, and rich APIs for seamless integration into business systems, all within a secure, isolated cloud network.

Foundation Models (FMs) and Customization

With access to powerful models from the Qwen open-source and Tongyi families, businesses can effortlessly tailor AI models using tools like SFT (Supervised Fine-Tuning) and LoRA (Low-Rank Approximation).

- **Built-In Workflows:** Transform model development with tools that facilitate model compression, inference acceleration, and evaluation through visual templates.
- **AI Agents:** Develop AI agents using an extensive set of plugins and visual orchestration services, making AI integration into enterprise systems straightforward and efficient.

Robust Security and Privacy

Model Studio ensures enterprise data confidentiality through comprehensive security measures, including:

- **Data Isolation:** Maintain separate environments for research and production, minimizing risks and ensuring data protection.
- **Content Security:** Automated content monitoring to detect and mitigate risks in generated content, maintaining ethical and legal compliance.
- **Network Protection:** Safeguard transmissions with PrivateLink and customizable network security policies, alongside defenses against DDoS and other attacks.

Scenarios and Use Cases

Seamless Agent Development

Utilize Model Studio's Assistant API for quick AI agent development, leveraging pre-configured plugins and SDKs that support mainstream programming languages, enabling efficient integration and deployment.

- **Prompt Engineering:** Access 160+ prompt templates with configurable variables for diverse scenarios, providing flexibility and customization.
- **Functionality Enhancement:** The platform's evolving architecture supports new algorithms, optimizing task execution and responses.

Model Development and Deployment

Model Studio offers an end-to-end solution for building and deploying customized models securely within your enterprise network:

- **Dataset Management:** Streamline data handling to refine models by adjusting parameters and deploying them with a single click.
- **Model Gallery:** Choose from a vast selection of models to suit various business needs and tailor them using advanced fine-tuning techniques.

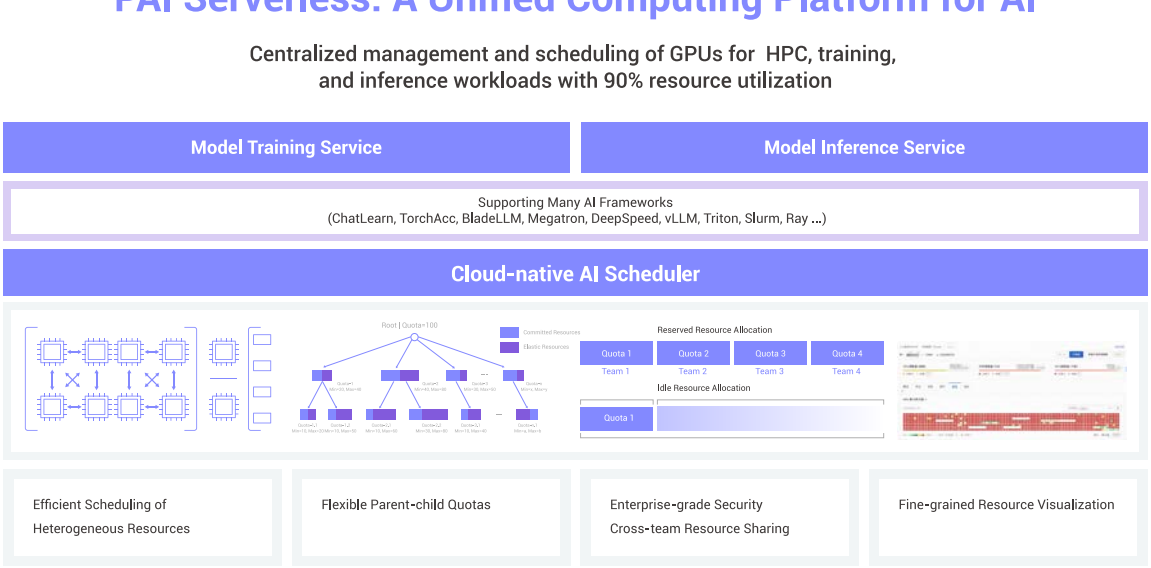
Alibaba Cloud Model Studio provides a robust, secure platform for businesses to develop and deploy generative AI, fostering digital innovation across industries. Explore tailored AI solutions with Alibaba Cloud.

2.3.2 PAI-EAS INFERENCE PLATFORM – ENABLING OPEN-SOURCE LLMS, RAG, AND AGENTS FOR GENERATIVE AI

Introduction

Alibaba Cloud's PAI-EAS platform significantly advances the accessibility of large language models and generative AI, fostering innovation and deployment in multiple sectors.

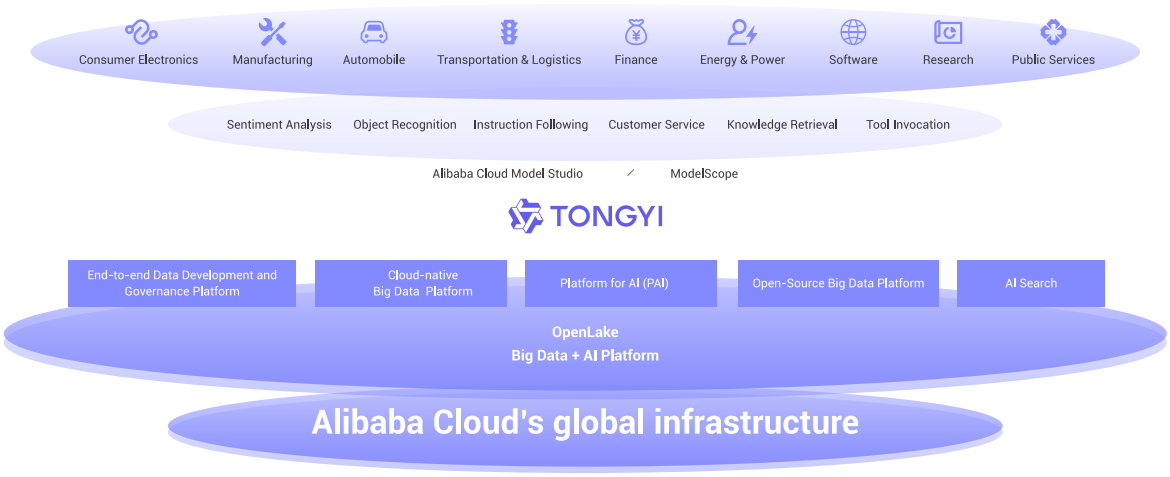
Overview of PAI-EAS Inference Platform



PAI-EAS is designed to handle the demanding requirements of modern generative AI applications. It offers:

1. **Scalable Deployment:** The platform can dynamically scale resources based on demand, ensuring optimal performance during peak loads.
2. **High Performance:** Leveraging advanced hardware like GPUs and optimized software stacks, PAI-EAS delivers low-latency inference with high throughput.
3. **Cost Efficiency:** Users benefit from pay-as-you-go pricing models, allowing them to manage costs effectively while accessing powerful computational resources.
4. **Ease of Use:** A user-friendly interface simplifies the deployment and management of complex models, making it accessible even to non-experts.

LLM-driven Innovations in Industrial AI Applications



Supporting Open-Source LLMs

PAI-EAS plays a crucial role in fostering an open-source ecosystem around LLMs. Key features include:

- **Wide Model Support:** The platform supports popular open-source models such as LLAMA2, Baichuan2, GLM3, Falcon, Mistral, and more. This broad support enables developers to choose the best model for their specific needs.
- **Pre-trained Models and Datasets:** Access to pre-trained state-of-the-art (SOTA) models and multiple open-source datasets lower the barrier to entry for developing custom applications.
- **Customization API Services:** Developers can quickly build customized industry models using just dozen of code lines since PAI-EAS provides half-ready API services.

Creating Retrieval-Augmented Generation (RAG) Systems

Retrieval-Augmented Generation (RAG) combines the strengths of retrieval-based and generative approaches, enhancing the quality and relevance of generated text. PAI-EAS facilitates the creation of RAG systems through:

- **Integrated Knowledge Bases:** Users can upload diverse data sources, including structured and unstructured documents, which are processed into vectorized forms optimized for fast retrieval.
- **Advanced Retrieval Algorithms:** Utilizing sophisticated algorithms, PAI-EAS efficiently retrieves relevant information from knowledge bases to augment the generation process.

- **Seamless Integration:** RAG capabilities are seamlessly integrated into the PAI-EAS workflow, allowing developers to leverage these features without additional complexity.

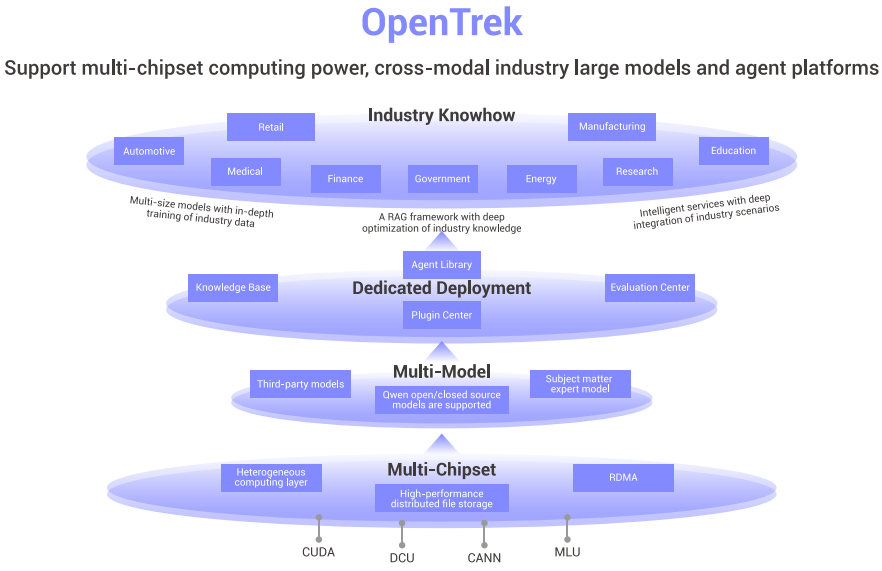
Empowering Intelligent Agents

Intelligent agents powered by PAI-EAS bring human-like interaction capabilities to various applications. Key functionalities include:

- **ChatLearn Framework:** An industry-level reinforcement learning from human feedback (RLHF) framework that supports supervised fine-tuning (SFT), reward modeling (RM), and RLHF pipelines. This framework ensures continuous improvement in agent performance.
- **BladeLLM:** Capable of handling sequences up to 280k tokens, BladeLLM provides unparalleled support for long-context conversations, making it ideal for building sophisticated chatbots and virtual assistants.
- **Flexible Resource Management:** Logical quotas enable efficient sharing of resources between training and inference tasks, maximizing utilization without compromising performance.

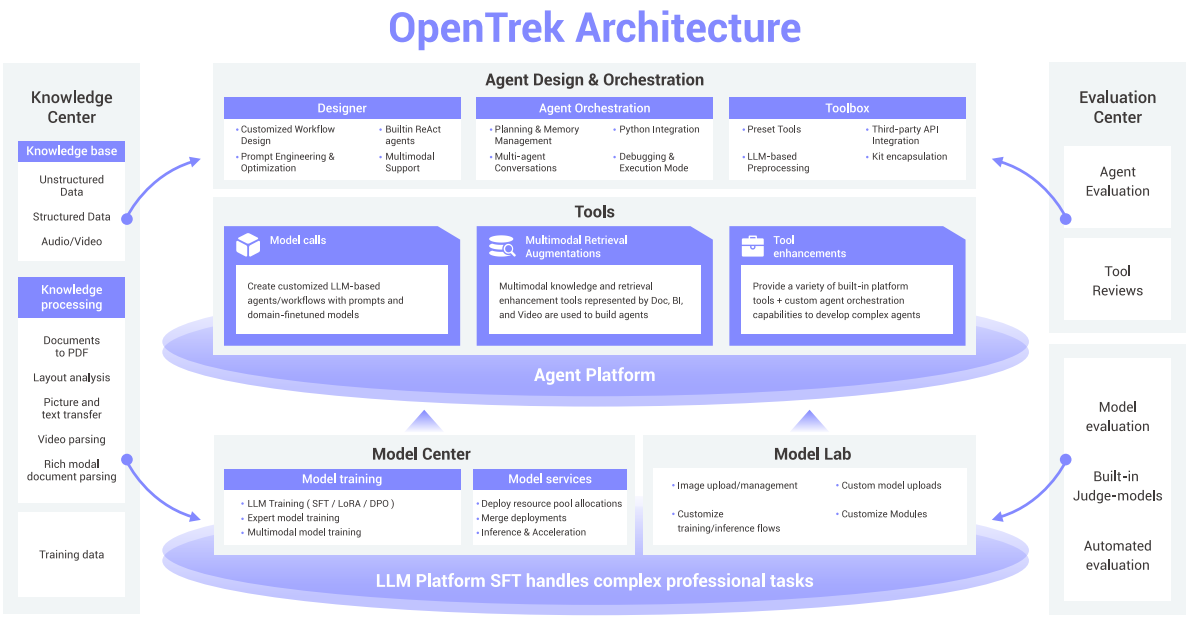
2.3.3 OPENTREK PLATFORM – ADVANCED GENERATIVE AI FOR INDUSTRY APPLICATIONS

Overview of OpenTrek



Opentrek is an enterprise-grade, all-inclusive platform designed for governmental and large-corporate clients to facilitate the application of AI services. It offers a comprehensive suite of functionalities, including private knowledge and corpus management, model training and inference, agent design and orchestrations, as well as agent evaluation and deployment services.

Key Features and Capabilities



- **High Reusability and Flexibility Framework:**
  - Adopts the OneRAG framework suitable for various scenarios such as document Q&A, data Q&A, and image-text understanding.
  - Offers flexible workflow orchestration capabilities, supporting the development of LLM agents for complex real-world scenarios.
- **Multimodal Knowledge Data Center with Efficient Data Parsing Capabilities:**
  - Supports parsing of over 29 file formats, including text, images, tables, formulas, etc.
  - Facilitates the construction of multimodal knowledge bases such as textual, tabular, and image-text knowledge bases.
- **Automated Evaluation System:**
  - Includes automated evaluation functions for large language models, expert models, and custom scenario machine evaluations, ensuring model quality and performance.
- **Unified Model Management:**
  - Can uniformly manage and accelerate various types of models, such as large language models, multimodal models, and expert models.
  - Supports importing third-party models for training and deployment within the platform, demonstrating strong extensibility.

- Enables the import of third-party models through the Model Laboratory, facilitating their training and deployment on the platform.

- **Enterprise-Level Multi-Tenant Resource and Algorithm Management:**
  - In a multi-tenant environment, utilizes a three-tier management system (clusters, resource pools, workspaces) to optimize the allocation of computing resources, improve utilization rates, and reduce resource idling.
  - Supports the sharing and reuse of models and algorithms across tenants and workspaces, providing one-click unified scheduling and operation under multiple resource pools to improve resource utilization efficiency.

- **Multimodal Agent and RAG Integration:**
  - Based on multi-modal image-text knowledge-enhanced retrieval, integrates additional plug-in tools and generative capabilities.
  - Utilizes multi-agents to achieve more comprehensive and rich knowledge acquisition and retrieval.

- **Model and Computing Optimization:**
  - Compatible with multiple heterogeneous GPUs, supports unified management and mixed inference scheduling.
  - Through joint optimizations in I/O scheduling, communication libraries, training and inference frameworks, and model quantization, significantly improves throughput, reduces latency, and achieves extreme acceleration in training and inference.

Flexible Deployment Options

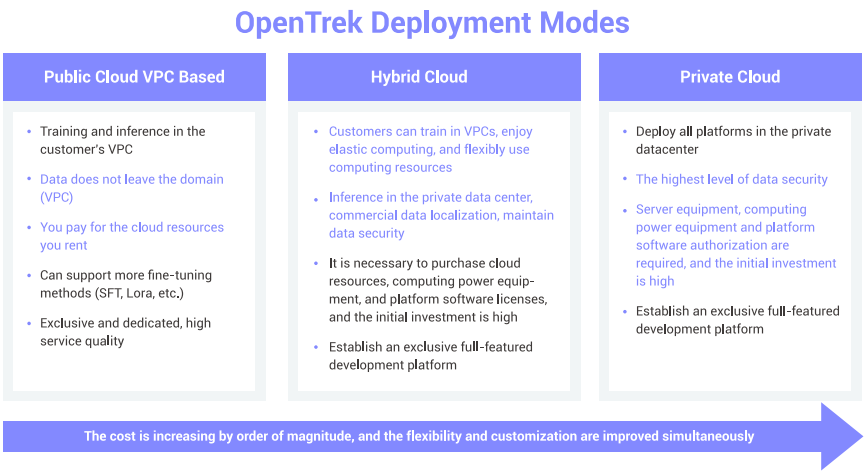
OpentTrek offers three deployment types: Public Cloud VPC, Hybrid Cloud, and Private Cloud, each suited to different needs.

Public Cloud VPC: Uses customer's VPC for training and inference, keeping data within the VPC. It's cost-effective, with pay-as-you-go cloud resources and support for advanced fine-tuning.

Hybrid Cloud: Balances VPC training with private datacenter inference, offering flexible resource use but requiring both cloud and on-premises investments.

Private Cloud: Fully controlled environment in a private datacenter, providing top data security and control, yet needing significant initial investment.

OpenTrek Deployment Modes



Future Outlook

As AI continues to advance, so too will the capabilities of OpenTrek. Alibaba Cloud remains committed to pushing the boundaries of what's possible across all industries. With ongoing research and development, OpenTrek aims to redefine how businesses operate and innovate, improving processes and outcomes worldwide.

For detailed information about OpenTrek and how it can benefit your organization, please refer to the official documentation and resources provided by Alibaba Cloud.

2.4 GENAI APPS - CHATBOT

The primary application scenario for large language models (LLMs) is chatbots, which serve as a key focus for leveraging enterprise knowledge bases to deliver intelligent Q&A. Alibaba has introduced several powerful AI services to help businesses quickly integrate and build Retrieval-Augmented Generation (RAG) systems, enabling them to rapidly harness the capabilities of LLMs.

These services include:

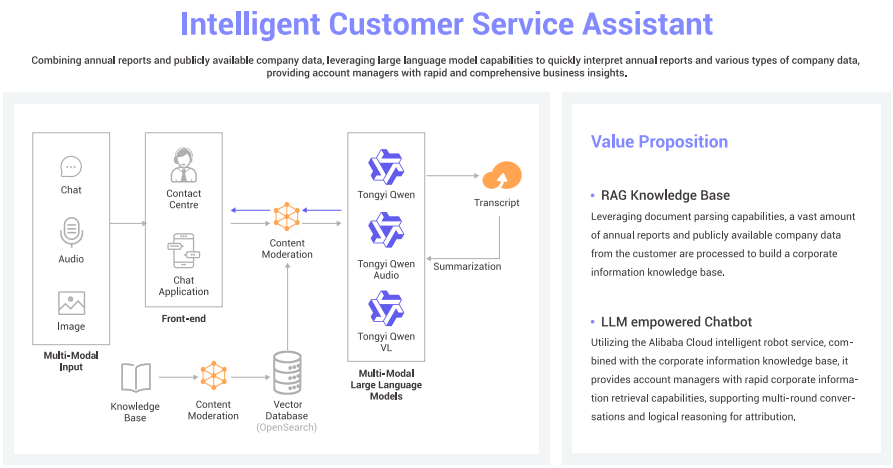
- **Alibaba Cloud Chatbot Service**
- **Quick Service**
- **AutoDoc by Opentrek**
- **Open Search**

Each of these services plays a crucial role in enabling businesses to swiftly implement intelligent chatbot solutions, enhancing customer interactions and operational efficiency.

2.4.1 ALIBABA CLOUD CHATBOT SERVICE

Alibaba Cloud Chatbot Service, based on Qwen LLM, provides 24/7 self-service consultation by deploying at enterprise customer service entry points such as websites, apps, and IM tools. The SaaS-based management system supports uploading various document formats (PDF, DOC/DOCX, TXT, Markdown, web pages, Excel, FAQs) and multi-turn dialogue flows as knowledge sources.

Offering comprehensive APIs, it enables enterprises to perform secondary integration and development, enhancing service experience and efficiency through intelligent capabilities like Retrieval-Augmented Generation (RAG) and large language models for powerful Q&A services.



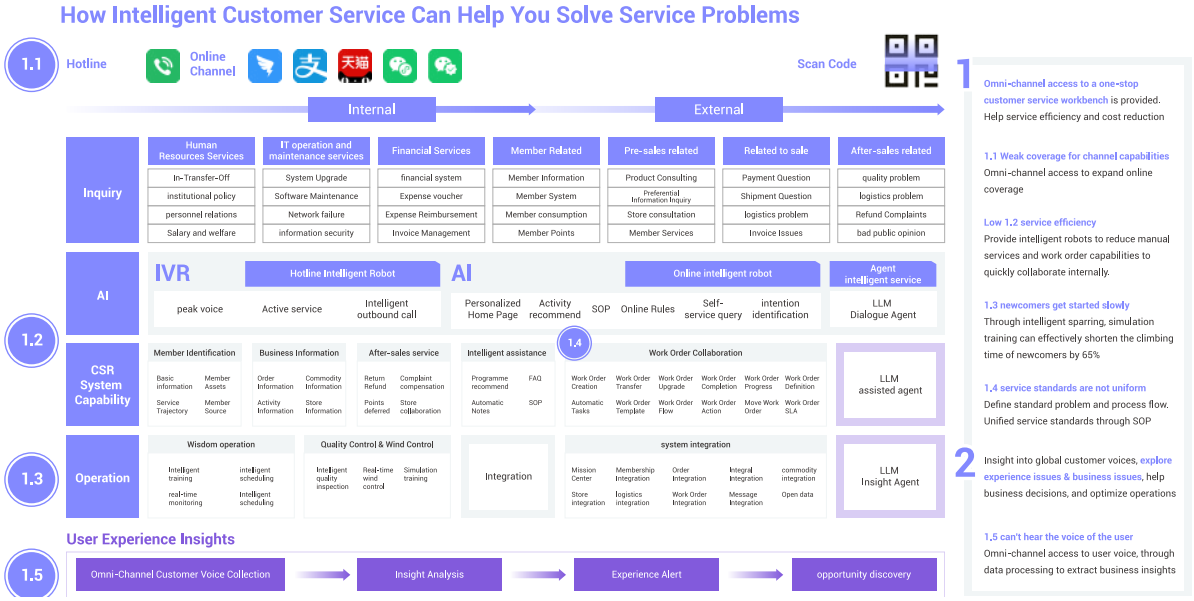
Key Advantages

- **Ready-to-Use :** Quick deployment without complex setup.
- **Operable and Reliable :** Flexible operational support ensuring accurate and reliable Q&A results.
- **Latest RAG Practices :** Utilizes cutting-edge RAG technology for state-of-the-art intelligent services.
- **Wide Industry Applications :** Successfully applied in insurance, government, e-commerce, retail, and more.
- **Domain-Specific Fine-Tuning :** Customized large models tailored to specific service domains for enhanced performance.

By leveraging these features and advantages, our product helps businesses quickly build efficient intelligent customer service systems, significantly improving customer satisfaction and operational efficiency across pre-sales inquiries, after-sales services, and daily customer interactions.

2.4.2 QUICK SERVICE

Quick Service, leveraging Alibaba's over 20 years of rich service operation experience, can help businesses quickly build a digital and intelligent customer service platform, significantly enhancing service quality. By integrating voice recognition, instant messaging, collaborative work orders, and advanced large language model capabilities, it not only helps businesses effectively reduce customer service costs but also enables them to respond swiftly to customer needs, thereby improving customer satisfaction.



Product Capabilities

- **Online Chat :** Provide service through online chat for external customers, quickly deployable on APP, Web, or WeChat ecosystem. Customer service agents respond to customer messages via the new retail online customer service console.
- **Hotline Support :** Offer telephone access for customers to reach customer service, with agents handling calls via the console.
- **Outbound Calling :** Agents can initiate outbound calls from the console to proactively serve external customers.
- **Work Order Management :** Record and track customer issues, enabling rapid collaboration among different customer service teams. Suitable for offline scenarios like applications and reviews.
- **Quality Inspection :** Ensure service quality by inspecting customer service interactions. The system provides real-time inspection, manual inspection, and AI-driven scoring.
- **Customer Service Assistant :** Enhance service efficiency and quality with a repository of effective solutions, helping agents quickly find answers to new or complex problems.

- **Service Insights :** Real-time monitoring of service traffic and quality, multi-dimensional analysis of service data, automatic generation of multiple service metrics, and intuitive visualization of service levels.

2.4.3 AUTODOC CHATBOT BY OPENTREK – ALIBABA CLOUD INTELLIGENCE'S ADVANCED CONVERSATIONAL AI SOLUTION

AutoDoc is a best practice tool for document-type RAG (Retrieval-Augmented Generation) applications based on the OpenTrek platform. It helps clients quickly achieve high-accuracy document question-answering capabilities based on large models.

Key Features and Advantages

- **Efficient Document Retrieval:** Seamlessly connects with user-created document knowledge bases on the OpenTrek platform, supporting fast parallel retrieval of a large number of documents in different formats.
- **Content Recognition and QA:** Capable of processing not only text content but also recognizing tables and images in documents, and referencing these elements when answering questions.
- **Flexible Configuration:** Provides an intuitive interface for configuring critical steps of the RAG process, such as data preparation, query planning, recall optimization, and summary generation.
- **Multiple Retrieval Modes:** Compatible with vector retrieval, keyword retrieval, and hybrid retrieval methods to meet different scenarios and needs.
- **Terminology Management:** Allows users to batch manage and define proprietary business terms and their explanations, enhancing answer specificity and accuracy.
- **Prompt Templates:** Includes preset general document QA prompt templates while also supporting custom adjustments to suit specific application scenarios.
- **Multi-turn Dialogue Support:** Ensures that the QA bot understands and continuously interacts with users, improving user experience.

- **Pre-trained Model Components:** Built-in model components fine-tuned by the OpenTrek platform, such as embedding models, reranking models, and generative large models, specifically optimized for document QA scenarios.
- **Real-time Debugging and Analysis:** Through the AutoDoc Laboratory, users can instantly test the performance of the document QA bot and optimize it based on feedback.
- **API Publishing:** Offers a convenient one-click publishing feature to provide document QA capabilities externally in the form of OpenAPI, facilitating system integration.

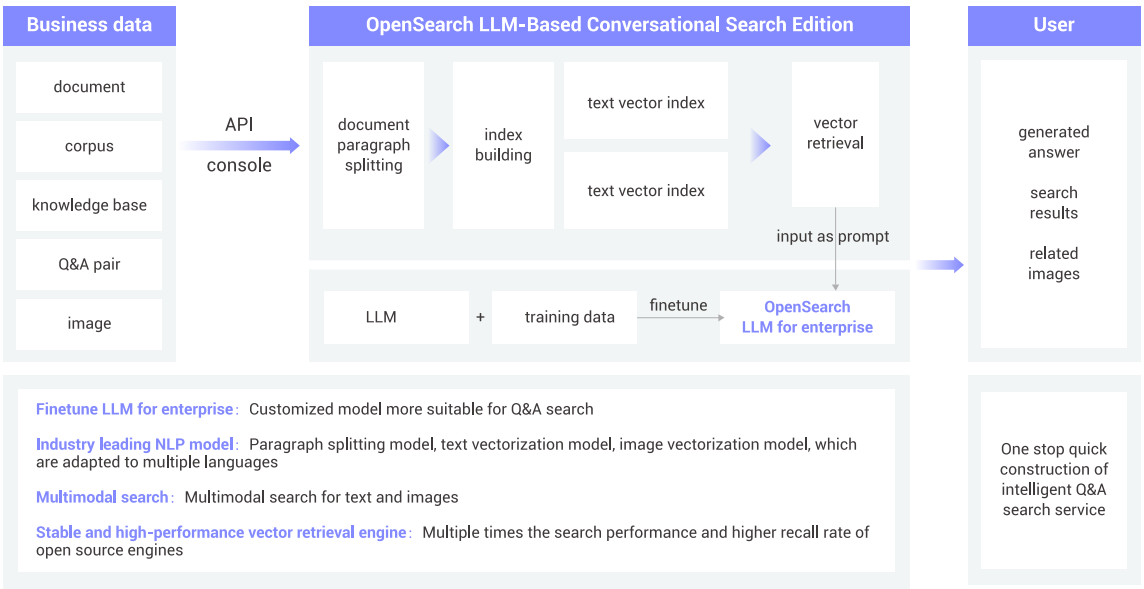
AutoDoc aims to simplify the process of building document QA solutions while ensuring flexibility and high performance, helping users make more effective use of their document resources.

2.4.4 OPENSEARCH

OpenSearch is a distributed search and analytics suite that makes it easy to perform and scale searches.

Advanced Features

- **Highly Scalable Search:** Capable of handling vast amounts of data with high-speed indexing and querying.
- **Real-Time Data Processing:** Supports real-time ingestion and processing of data, ensuring up-to-date information.
- **Comprehensive Analytics:** Offers robust analytics tools for deeper insights into data trends and patterns.
- **User-Friendly Interface:** Features an intuitive interface and API for seamless integration with existing workflows.



Creating a RAG system involves two main components: retrieving relevant documents from a knowledge base and generating coherent responses based on retrieved information. OpenSearch excels in the retrieval phase by providing:

1. **Efficient Document Indexing:** Quickly indexes and organizes unstructured data, making it readily accessible for retrieval.
2. **Advanced Query Capabilities:** Supports complex queries, including full-text search, fuzzy matching, and faceted navigation.
3. **Scalable Infrastructure:** Handles large-scale datasets efficiently, ensuring fast and accurate retrieval even under heavy loads.

2.4.5 CONCLUSION

Organizations across various industries have successfully implemented RAG-based chatbot solutions on Alibaba Cloud. These implementations highlight the versatility and adaptability of RAG-based chatbot solutions, making them suitable for any sector.

Embrace the future of intelligent customer engagement with RAG-based chatbot solutions on Alibaba Cloud. Unlock new possibilities for your business, regardless of your industry, and transform how you interact with your customers.

Use Cases and Applications

Industry	Use Cases and Scenarios
Retail	• Automate customer inquiries about product availability, pricing, and promotions. • Provide personalized recommendations based on customer history.
Finance	• Streamline loan application processing and compliance checks. • Offer real-time financial advice and portfolio analysis.
Healthcare	• Assist in patient triage and appointment scheduling. • Provide health-related information and answer FAQs.
Manufacturing	• Monitor production metrics and forecast demand. • Optimize supply chain operations by tracking inventory levels and delivery schedules.
Government and Public Sector	• Monitor production metrics and forecast demand. • Optimize supply chain operations by tracking inventory levels and delivery schedules.
Education	• Answer student inquiries about course offerings, schedules, and academic policies. • Provide support for administrative tasks like enrollment and fees.
Travel and Hospitality	• Handle booking inquiries and manage reservations. • Provide travel tips and local recommendations based on customer preferences.

Industry	Use Cases and Scenarios
Telecommunications	• Resolve common technical issues and troubleshoot problems. • Manage customer accounts and billing inquiries.
More Industries	• Generic Knowlegebase QnA

2.5 GENAI APPS - DOCUMENT INTELLIGENCE

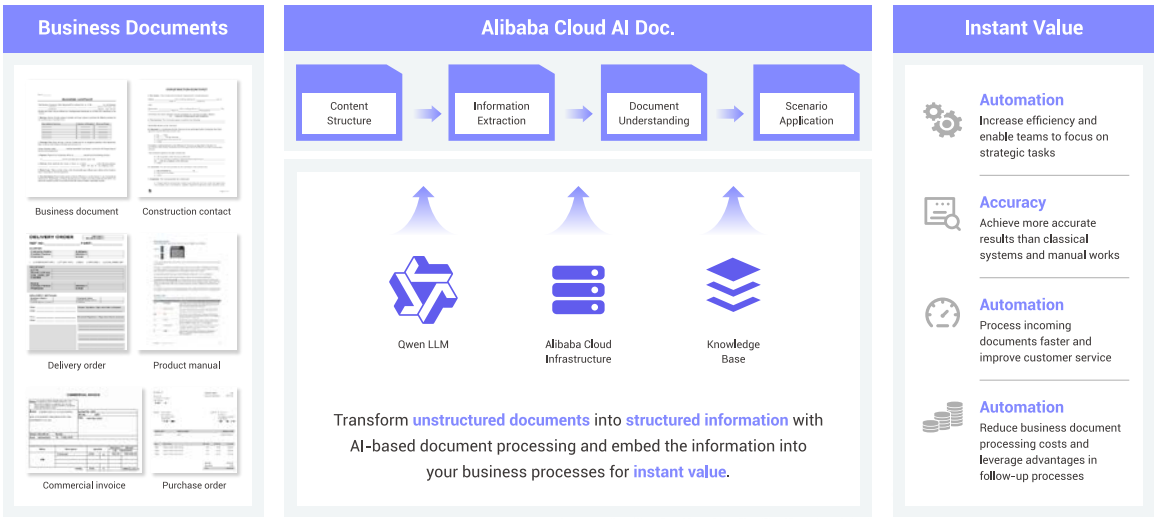
AIDoc, part of Alibaba Cloud's comprehensive suite of generative AI solutions, is designed to transform document processing and intelligence.

Introduction

AIDoc is a versatile tool that seamlessly integrates into existing workflows, offering key capabilities for document intelligence:

- **Document Understanding :** Extracts insights from unstructured text, tables, and images.
- **Content Generation :** Generates high-quality, contextually relevant content.
- **Automation :** Automates repetitive document tasks, reducing manual effort and errors.
- **Customization :** Supports API and SDK integration for tailored solutions.

Alibaba Cloud AI Document Processing Solution



Key Features

Category	Feature	Description
Intelligent Document Processing	Optical Character Recognition (OCR)	Converts scanned documents to editable data.
	Natural Language Understanding (NLU)	Extracts key information from document text.
	Data Extraction	Automatically extracts structured data.
Advanced Content Generation	Summarization	Generates concise summaries of documents.
	Translation	Translates documents while preserving context.
	Template-Based Generation	Generates standardized documents using templates.
Automation and Workflow Integration	Rule-Based Processing	Automates decisions based on document content.
	API Integration	Integrates with existing systems via APIs.
	Batch Processing	Handles large volumes of documents efficiently.
Security and Compliance	Data Encryption	Protects sensitive information during storage and transmission.
	Access Control	Manages permissions with role-based access.
	Audit Trails	Maintains logs for auditing and compliance.

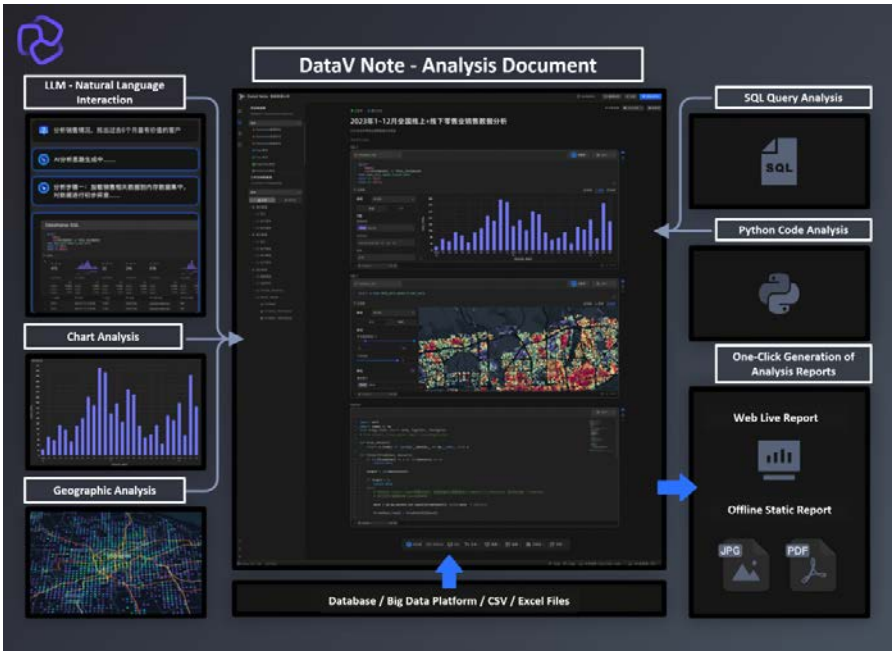
Use Cases and Applications

Industry	Use Cases
Legal Industry	Automates contract review and drafting, reducing time and costs while improving accuracy.
Financial Services	Streamlines loan application processing and compliance checks, enhancing operational efficiency.
Healthcare	Assists in medical record management and patient data analysis, supporting better clinical decisions.
Government and Public Sector	Enhances document management and reporting processes, promoting transparency and accountability.

2.6 GENAI APPS - BI ANALYSIS

2.6.1 DATAV NOTE – EMPOWERING INTERACTIVE DATA INSIGHTS ON ALIBABA CLOUD

Alibaba Cloud's DataV platform, enhanced with advanced Large Language Model (LLM) capabilities, introduces a groundbreaking approach to data analysis.



The DataV-Note (Intelligent Analysis) feature offers the following benefits:

- **AI-Driven Smart Analysis :** Harness the power of large models to automatically plan and execute data analysis tasks. With a single click, generate comprehensive analysis reports that cover the entire workflow—from data extraction and analysis to presentation and insights.
- **Diverse Analysis Methods :** Whether you are a coding expert or a business-side data enthusiast, DataV-Note provides a versatile suite of analysis methods. These include AI-driven analysis, visual analysis, SQL analysis, and Python analysis—all seamlessly integrated into one unified platform.
- **Collaborative Multi-User Analysis :** Facilitate real-time collaboration among data scientists, data engineers, and business experts on a single platform. Share insights and analyze data together effortlessly, enhancing teamwork and productivity.
- **One-Click Report Generation :** Utilize an intuitive, document-style interface that naturally integrates the data analysis process with its results. Eliminate the need for secondary editing by easily creating, editing, and publishing analysis reports with minimal effort.

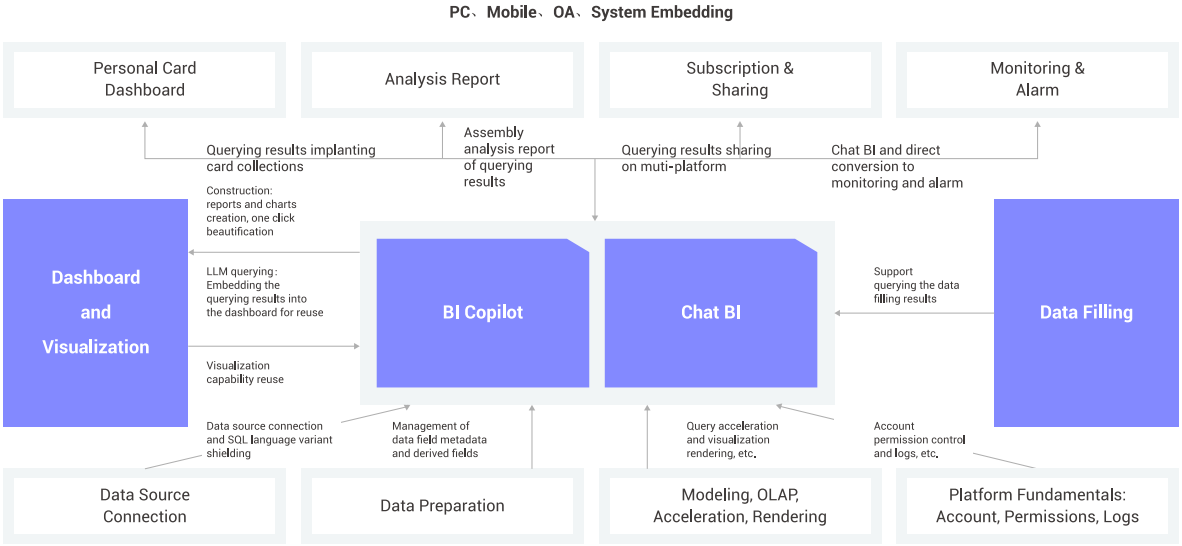
By enabling intuitive, conversational interactions through ChatBI, Alibaba Cloud empowers organizations to unlock the full potential of their data assets, yielding immediate and actionable insights. This powerful combination sets new standards for data visualization and analysis, driving smarter and faster decision-making across industries. As part of Alibaba Cloud's commitment to innovation, DataV-Note is at the forefront of this revolution, ensuring businesses can harness the power of their data more effectively than ever before.

2.6.2 QUICK BI – REVOLUTIONIZING DATA INSIGHTS ON ALIBABA CLOUD

In the rapidly evolving landscape of business intelligence (BI), the demand for intuitive and interactive data exploration tools has never been higher. Alibaba Cloud's Quick BI platform, augmented with advanced ChatBI capabilities, offers a revolutionary approach to data analysis.

ChatBI represents a significant leap forward in business intelligence by integrating natural language processing (NLP) and generative AI into data analysis workflows. Users can now interact with Quick BI using conversational interfaces, enabling them to request specific insights or visualizations simply by sending messages. This capability not only enhances accessibility but also streamlines the process of data exploration, allowing users to obtain immediate and actionable insights.

The Relation Between Smart Q and Other Functions of Quick BI



Key Features of ChatBI with Quick BI

1. **Natural Language Interaction :** Users can ask plain-language questions and receive instant, accurate, chart-based responses powered by advanced NLP algorithms.
2. **Real-Time Data Visualization :** Dynamically generates up-to-date charts and graphs in various types (bar, line, pie, heat maps) tailored to different analytical needs.
3. **Interactive Exploration :** Refine queries iteratively, drill down into specific data points, and interactively filter, sort, and manipulate data within charts for deeper analysis.

4. **Contextual Awareness :** Maintains conversation context for seamless multi-turn dialogues and dynamically adjusts visualizations to provide deeper insights.
5. **Automated Insights Generation :** Provides narrative summaries, key takeaways, and data-driven recommendations to help users quickly grasp trends and make informed decisions.
6. **Seamless Integration :** Integrates with Quick BI dashboards and diverse data sources, incorporating ChatBI-generated insights into comprehensive analytical frameworks.

Benefits of ChatBI with Quick BI

Benefit	Description
Enhanced Accessibility	Simplifies data analysis for non-technical users, enabling everyone in the organization to derive value from data.
Increased Efficiency	Reduces the time and effort required to generate insights, allowing users to focus on strategic decisions rather than data wrangling.
Improved Accuracy	Leverages advanced NLP and AI models to ensure precise interpretation of user queries and accurate generation of visualizations.
Faster Decision-Making	Provides real-time insights that empower users to make informed decisions quickly, staying agile in dynamic business environments.

Use Cases and Applications

Industry	Use Cases and Applications
Retail	Retailers can monitor sales performance, track inventory levels, and identify top-selling products by region or category. Example query: "What were the top 5 selling items in North America last month?" (Response: Bar chart)
Finance	Financial institutions can analyze trading volumes, evaluate investment portfolios, and detect market trends. Example query: "Display the stock performance of our top clients over the past quarter." (Response: Line chart)
Healthcare	Healthcare providers can assess patient outcomes, monitor treatment efficacy, and optimize resource allocation. Example query: "Show the number of patients treated by department in the last week." (Response: Pie chart)
Manufacturing	Manufacturers can track production metrics, forecast demand, and streamline supply chain operations. Example query: "Generate a heatmap showing machine uptime across different factories." (Response: Heatmap)

ChatBI capabilities integrated with Alibaba Cloud's Quick BI platform represent a transformative shift in how businesses engage with their data. By enabling intuitive, conversational interactions that yield immediate and actionable insights, ChatBI empowers organizations to unlock the full potential of their data assets. As part of Alibaba Cloud's commitment to innovation, this powerful combination sets new standards for data visualization and analysis, driving smarter and faster decision-making across industries.

2.6.3 AUTOBI WITH OPENTREK

The AutoBI tool on the OpenTrek platform aims to streamline the building of NL2SQL services while ensuring flexibility and high performance.

Key Features and Advantages

- **Proprietary Business Terminology and Metric Calculation:**
  - o Extends the understanding of specialized terms in specific business domains.
  - o Supports calculations of business metrics ranging from simple to complex, including multi-table calculations and multi-record single-table calculations.
- **Optimized Models and Algorithms for NL2SQL Scenarios:**
  - o Includes features such as multi-turn conversation support, business terminology understanding, and schema information retrivial to improve answer accuracy and efficiency through ReAct and NL2SQL technologies.
  - o Built-in pre-trained model components optimized for BI Q&A scenarios, such as embedding, ranking models, and Text2SQL large models.
- **Support for Temporal Diversity:**

Supports independent time, colloquial time, and continuous time descriptions, suitable for various time expression methods.

- **Intelligent Structured Database Support:**
  - o Supports one-click integration with mainstream structured databases and automatic metadata retrieval.
  - o Flexible Configuration and Retrieval Modes:
  - o Offers an intuitive interface to configure key steps of the NL2SQL RAG process.
  - o Compatible with vector retrieval, keyword retrieval, and hybrid retrieval to meet different needs.
- **API Publishing Feature:**
  - o Provides one-click publishing of OpenAPI of NL2SQL services for easy system integration.

Through the above features, AutoBI is committed to providing users with an efficient and flexible NL2SQL solution, enabling users to query their databases without any technical barriers.

2.7 GENAI APPS - SPEECH

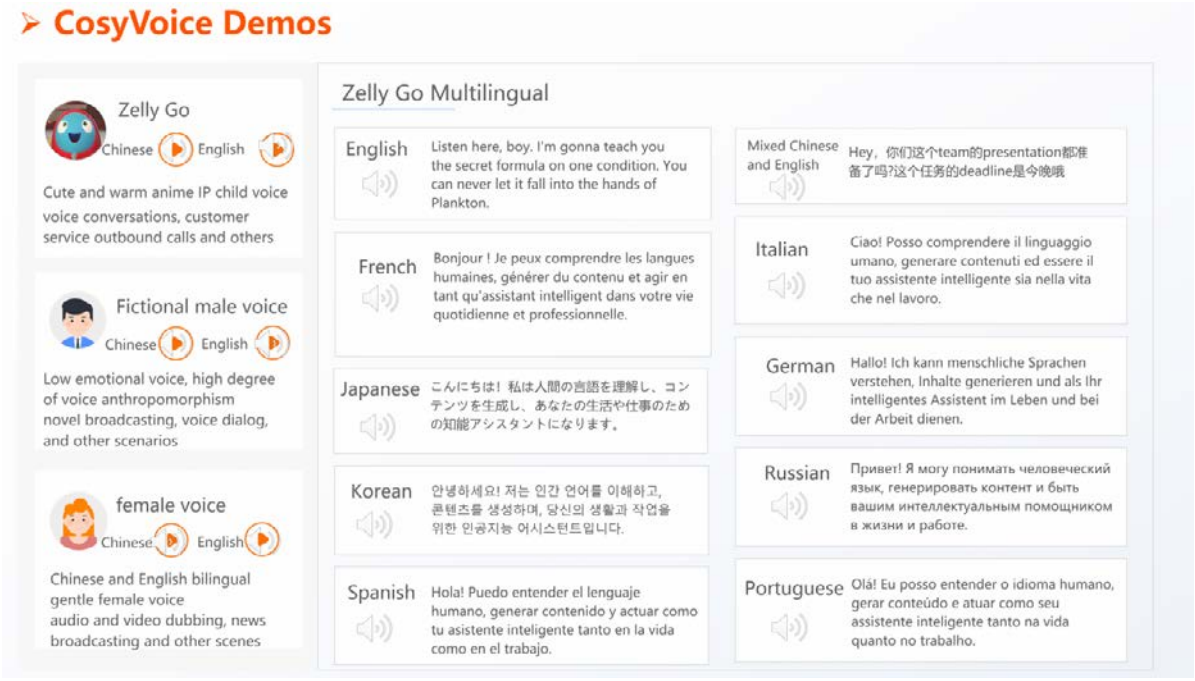
In an interconnected world, the demand for advanced audio processing technologies is growing. Alibaba Cloud addresses this need with GenAI-Audio , a suite of tools including Automatic Speech Recognition (ASR), Text-to-Speech (TTS), intelligent speech interaction, and innovative applications like Qwen-Audio . These technologies transform audio data into meaningful interactions and insights using deep learning and large datasets.

GenAI-Audio supports multiple languages and offers high-accuracy ASR and TTS functionalities. Key models include Paraformer and SenseVoice for ASR, and Cosy Voice and Sambert for TTS. Alibaba Cloud's Intelligent Speech Interaction services facilitate real-time and file-based ASR, enhancing applications like customer service bots and virtual assistants. Qwen-Audio further improves voice interactions with advanced features such as role separation and speaker diarization.

By integrating these capabilities, businesses and developers gain access to comprehensive solutions that enhance customer service, automate workflows, and create immersive entertainment experiences, showcasing Alibaba Cloud's commitment to AI innovation.

2.7.1 COSYVOICE – LEADING TEXT-TO-SPEECH SOLUTION

Alibaba's CosyVoice is a cutting-edge Text-to-Speech (TTS) solution that transforms written text into highly realistic and natural-sounding speech. Using advanced neural network models and deep learning algorithms, CosyVoice achieves exceptional fidelity and expressiveness. It supports over 28 voice styles, including general, customer service, education, digital conversation, and regional dialects, catering to diverse use cases and industries.



Multilingual speech generation

Key Features of CosyVoice

- **Ultra-Human-Like Experience :** Achieves up to 98% similarity to real human voices with advanced product-level optimizations for high-quality, professional audio output.
- **Natural Listening Experience :** Trained on vast audio datasets, CosyVoice delivers rich, dynamic, and emotionally expressive speech with a Mean Opinion Score (MOS) exceeding 4.5, surpassing industry standards.
- **Diverse Voice Styles :** Supports over 28 distinct voice styles tailored for various scenarios like customer service, chatbots, education, and media dubbing. Includes specialized voices such as "Jelly Bean," a warm child-like voice, and "Novel Male Voice," known for its deep and emotional tone.
- **Multilingual Support :** Excels in multilingual synthesis, supporting languages like Chinese, English, Cantonese, Japanese, Korean, Spanish, French, German, and many others. Offers seamless language switching within sentences or contexts, enhancing global versatility.
- **Customizable Voice Replication :** Utilizes advanced feature extraction to replicate individual voices with up to 95% similarity in seconds, providing rapid and cost-effective solutions for personalized voice creation suitable for both ordinary users and professionals.

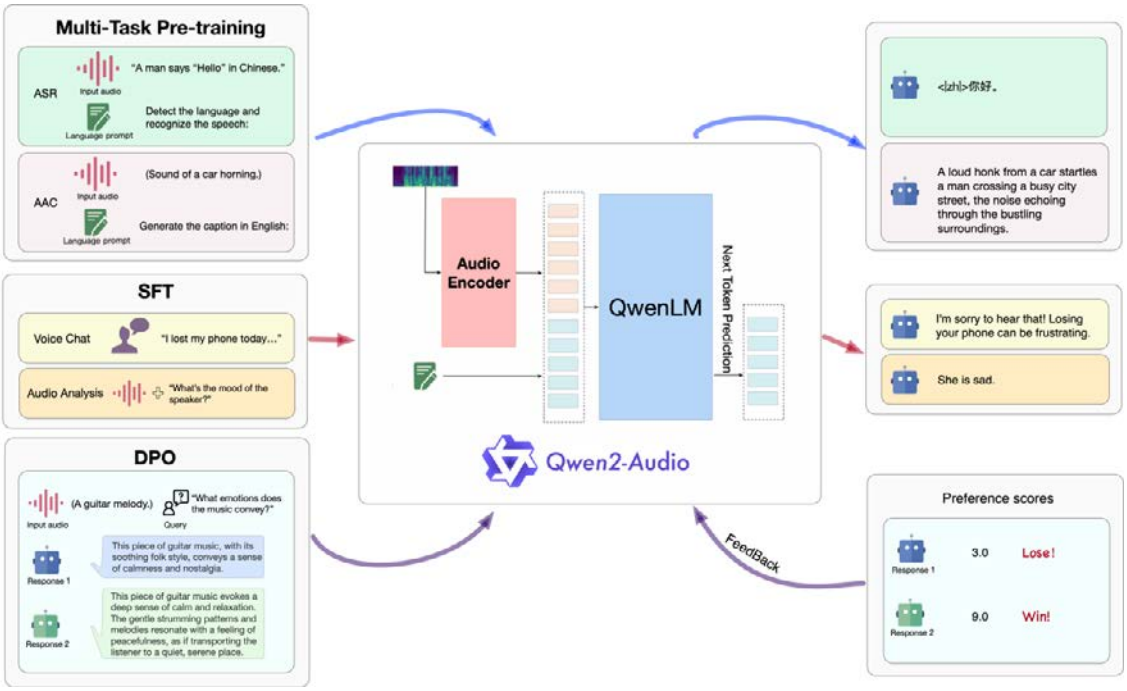
- **Seamless Integration :** Provides SDKs for Java, Python, and client-side applications, ensuring easy integration into various business environments and compatibility with existing systems and workflows.

Use Cases and Applications

Industry	Use Case
Customer Service	Automate call centers with highly realistic voices that enhance user experience and reduce operational costs.
Entertainment and Media	Provide immersive experiences through high-quality voiceovers for movies, games, and audiobooks.
Education	Create interactive learning materials with engaging narrations that captivate students' attention.
Smart Devices	Enhance IoT devices with conversational interfaces that respond naturally to user commands.
Accessibility	Offer text-to-speech functionalities that assist visually impaired individuals in accessing information.

2.7.2 QWEN-AUDIO – ADVANCED MULTIMODAL AUDIO LANGUAGE MODEL

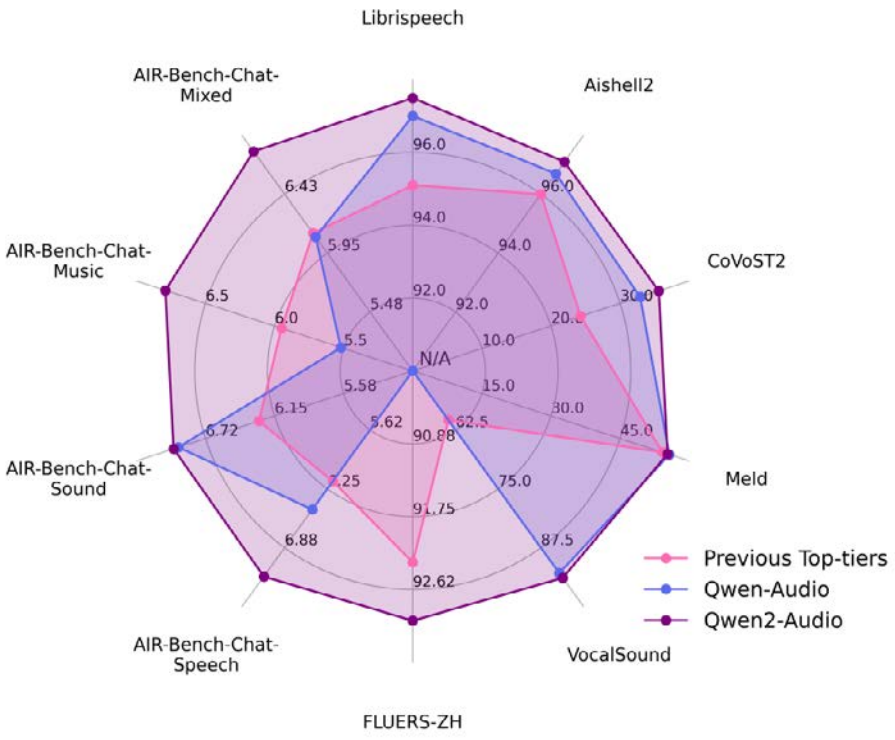
Alibaba's Qwen-Audio is a groundbreaking multimodal large audio language model that integrates advanced speech recognition, understanding, and generation capabilities. By leveraging a multi-task training framework and state-of-the-art algorithms, Qwen-Audio addresses the challenge of varying textual labels across datasets, enabling robust knowledge sharing and handling diverse audio tasks without task-specific fine-tuning. Incorporating over 30 tasks, it has demonstrated strong performance across multiple benchmark evaluations.



Training architecture of Qwen-Audio

Key Features

- **Multi-Task Learning Framework :** Manages textual label variations and facilitates knowledge sharing for better generalization across diverse audio data.
- **State-of-the-Art Performance :** Achieves top rankings in Acoustic Scene Classification (CochlScene, TUT 2017), Audio Classification (VocalSound), and strong performance in Speech Recognition (LibriSpeech) and Emotion Recognition (MELD).
- **Flexible Multi-Run Chat :** Supports multi-turn dialogues, real-time responses, and creative capabilities for enhanced user experience.
- **Advanced Models :** Combines Qwen-7B (LLM) and Whisper-large-v2 (audio encoder) for multi-task audio understanding; Qwen-Audio-Chat adds multimodal alignment for flexible interactions.
- **Extensive Evaluation :** Demonstrates superior performance on 12 standard benchmarks across various audio processing tasks.

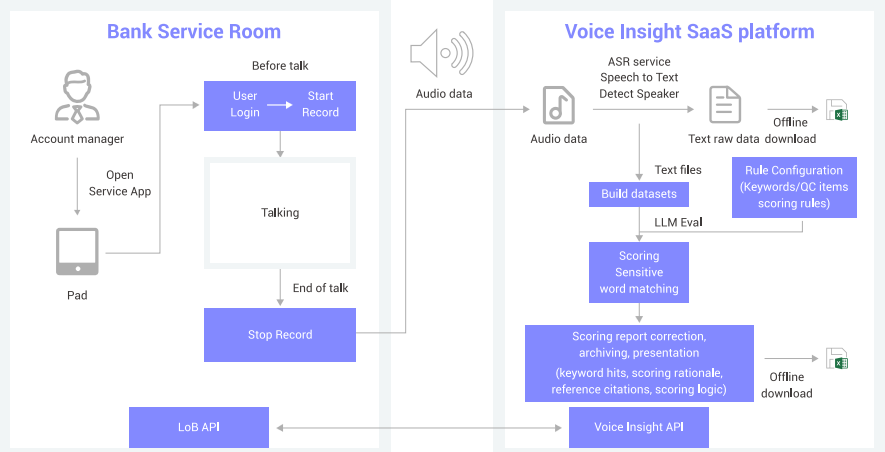


Qwen2-Audio significantly surpasses previous SOTAs

Industry	Use Case
Customer Service	Automate call centers with highly accurate speech recognition and natural language understanding, enhancing customer interactions.
Media and Entertainment	Provide immersive experiences through high-quality voiceovers for movies, games, and audiobooks.
Education	Create interactive learning materials with engaging narrations that captivate students' attention.
Healthcare	Assist healthcare providers by accurately transcribing patient consultations and generating summaries for medical records.
Smart Devices	Enhance IoT devices with conversational interfaces that respond naturally to user commands.
Accessibility	Offer text-to-speech functionalities that assist visually impaired individuals in accessing information.

2.7.3 VOICE INSIGHT – SOLUTION FOR ENSURING SUPERIOR CUSTOMER SERVICE

Alibaba's Voice Insight solution leverages advanced generative AI, integrating ASR, NLP, and ML to automate and enhance call center evaluations. By analyzing audio data, Voice Insight identifies agent performance improvements, policy compliance, and overall interaction quality, providing accurate transcriptions, contextual understanding, and issue flagging. This streamlined approach not only ensures high standards in call center operations but also offers valuable insights for training and development.



Key Features of Voice Insight

- **Automatic Speech Recognition (ASR) :** Converts spoken language into text with high accuracy using state-of-the-art models; supports multiple languages for global applicability.
- **Natural Language Processing (NLP) :** Analyzes transcriptions to identify sentiment, intent, and compliance; detects nuances indicating dissatisfaction or areas needing attention.

- **Machine Learning (ML) Models :** Integrates ML models trained on vast datasets to predict outcomes and identify patterns; customizable models tailored to specific industry requirements.
- **Real-Time Monitoring and Feedback :** Provides real-time monitoring for supervisor intervention during live calls; offers instant feedback to agents for skill improvement and adherence to best practices.
- **Comprehensive Reporting and Analytics :** Generates detailed reports on call metrics, trends, and agent performance; facilitates data-driven decision-making with actionable insights.
- **Integration with Existing Systems :** Seamlessly integrates with CRM, contact center, and workforce management systems; centralizes quality assurance processes for enhanced operational efficiency.

Use Cases and Applications

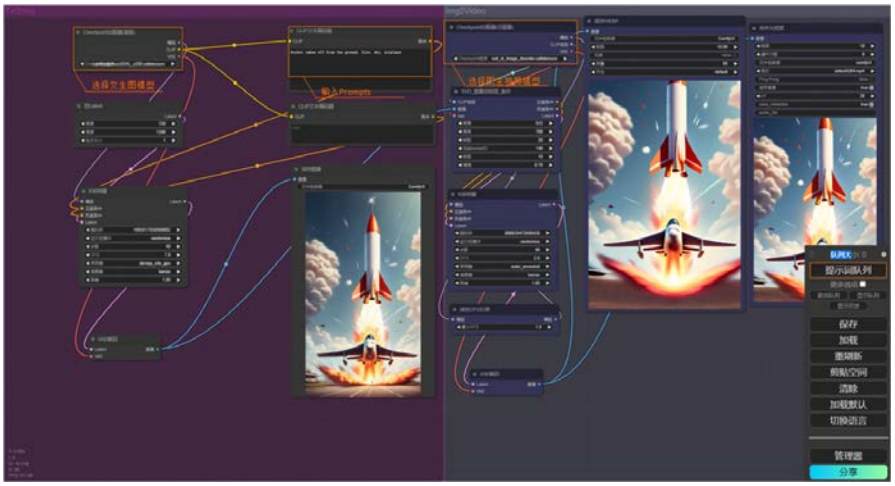
Industry	Application Scenario
Retail	Global Retail Chain:Achieved a 20% increase in first-call resolution rates after adopting Voice Insight for real-time coaching and feedback.
Financial Services	Financial Services Firm:Reduced compliance violations by 30% through automated policy adherence checks during customer interactions.
Telecommunications	Telecommunications Provider:Enhanced agent engagement and retention by providing personalized development plans based on Voice Insight analytics.

2.8 GENAI APPS - VISION

2.8.1 PAI-EAS - TEXT2IMAGE AND COMFYUI

PAI-EAS is an elastic algorithm service that simplifies the process of deploying AI models into production environments. It offers robust features such as automatic scaling, traffic management, and monitoring, making it ideal for handling varying workloads efficiently. By leveraging PAI-EAS, developers can deploy models with ease, ensuring high performance and reliability.

COMFYUI is a user-friendly interface integrated into PAI-EAS, designed to enhance the usability and accessibility of generative AI models. It offers an intuitive experience for both novice and experienced users, empowering them to leverage complex AI technologies effortlessly.



Key Features

1. **Simplified Workflow:** Guides users through the model deployment and inference process with clear instructions and visual aids, minimizing learning curves.
2. **Interactive Controls:** Provides interactive sliders, dropdown menus, and input fields, allowing users to adjust parameters in real-time and see immediate results.
3. **Real-Time Feedback:** Offers instant feedback on generated outputs, helping users refine their inputs and achieve desired outcomes more quickly.
4. **Customizable Templates:** Includes pre-built templates for common use cases, which can be customized to meet specific requirements.
5. **Collaboration Tools:** Supports collaboration by enabling multiple users to work together on projects, share settings, and review results collaboratively.

Use Cases and Applications

Industry	Application Scenario
Advertising	Marketing Agency:Leveraged Text2Image capabilities to automate the creation of visually appealing social media posts, significantly reducing production time and costs.
Product Design	Product Design Firm:Utilized Comfy UI to streamline the design prototyping process, enabling faster iterations and more accurate representations of final products.
Creative Services	Creative Studio:Employed both Text2Image and Comfy UI to explore new artistic directions, producing unique and engaging visual content for various projects.
Gaming	Game Development Company:Used Comfy UI to create detailed and interactive prototypes for game environments, enhancing the design and development phases.
Media & News	News Organization:Applied Text2Image to generate dynamic visuals for news articles and reports, improving reader engagement and content richness.

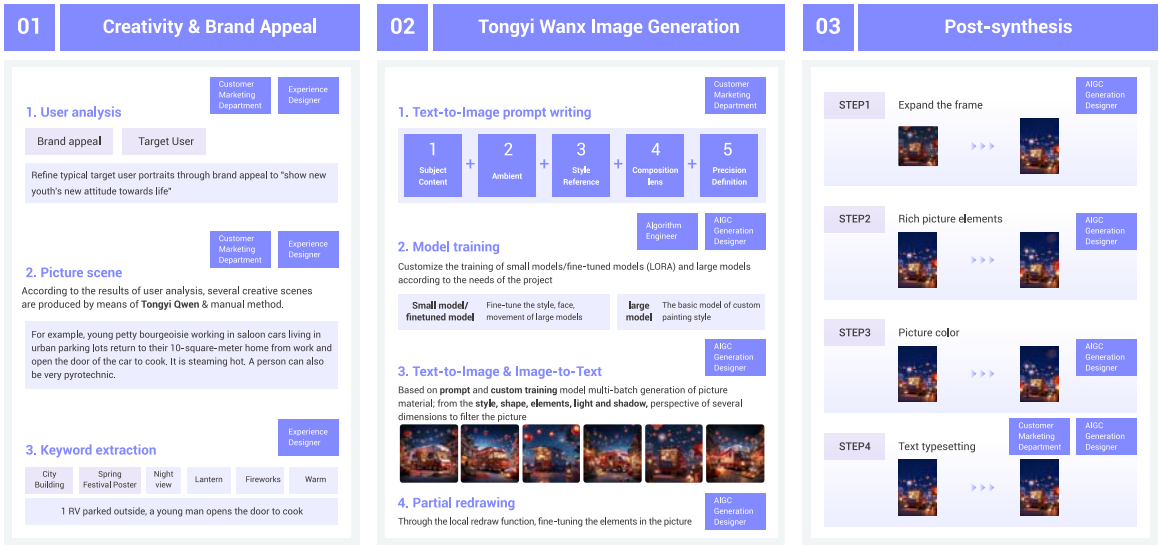
2.8.2 ADVERTISEMENT GENERATION – LEVERAGING GENERATIVE AI FOR MARKETING EXCELLENCE

In the era of digital transformation, **creating engaging and personalized advertisements** is crucial for brands to capture consumer attention. **Alibaba Cloud's generative AI (AIGC) technologies** revolutionize advertising by:

- **Automating content creation** , enabling businesses to generate high-quality social media content such as posters, advertisements, images, and videos efficiently.
- **Enhancing personalization** through advanced algorithms that ensure precise control over the generation process, aligning outputs with brand guidelines and marketing objectives.
- **Optimizing marketing strategies** by facilitating interactive campaigns that foster deeper consumer connections and drive higher conversion rates.

These capabilities reduce costs , **improve productivity** , and provide innovative solutions for marketers to meet the demands of today's dynamic market.

Image Design Delivery Service with Tongyi Wanx



Use Cases and Applications

Feature	Use Case
Customized New Year Wishes and Interactive Campaigns	Consumers can scan QR codes to upload wishes and photos, customizing New Year greetings with personalized videos and voices. This enhances user engagement and creates memorable, shareable content.
AI Co-Creation Events	Users interact with AIGC to personalize New Year paintings, collaborating with celebrities and popular IPs. This interactive approach attracts fans and promotes brand loyalty.
Dynamic Posters and Packaging Design	AIGC technology integrates symbols into scenic environments, enhancing brand visibility and making advertisements more impactful. Scene marketing techniques ensure visually appealing presentations.
E-commerce Banner Layout	AI-driven layout design optimizes e-commerce banners, generating background images, virtual models, and try-on features for fashion items. This ensures visually engaging and effective banner designs.
Social Media Engagement and Content Creation	AIGC generates personalized creative photography and portraits, quickly gaining popularity on social platforms. Convenient operation via WeChat Mini Programs and a social growing mechanism lower user barriers and accelerate adoption.
Virtual Models and Background Replacement	AIGC enables the creation of virtual models and seamless background replacements, allowing brands to present products in diverse settings without physical shoots. High-definition photo restoration ensures professional-quality images.
Scene-Based Marketing and Interactive Experiences	During festivals, AI writes couplets based on user-specified themes, engaging users through mini-programs. LBS positioning technology adds a personal touch by locating the user's hometown and integrating local landmarks into the experience.

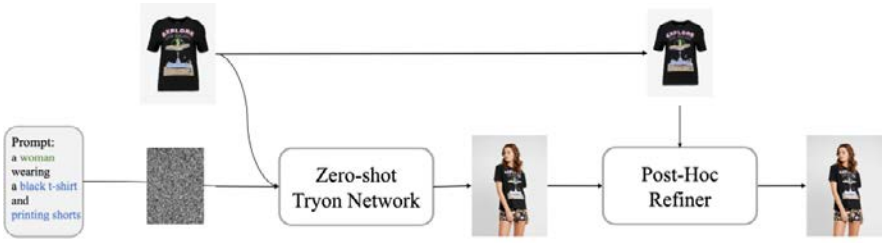
Advantages of AIGC in Advertising

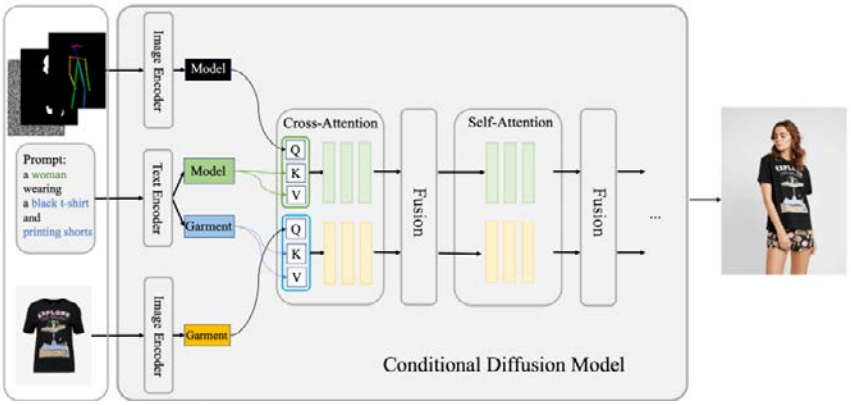
Alibaba Cloud's AIGC solutions offer several advantages over traditional advertising methods:

- **Personalization:** AIGC allows for highly customized content tailored to individual preferences, increasing relevance and engagement.
- **Efficiency:** Automated generation processes significantly reduce the time and resources required for content creation, enabling faster turnaround times.
- **Scalability:** AIGC can handle large volumes of content generation without compromising quality, supporting scalable marketing campaigns.
- **Cost-Effectiveness:** By automating repetitive tasks, AIGC reduces labor costs associated with manual content creation.
- **Creativity:** AIGC introduces new possibilities for creative expression, pushing the boundaries of what is possible in advertising.

2.8.3 VIRTUAL TRY-ON – HIGH-QUALITY MAKEUP TRANSFER MODEL

The conditional Diffusion Model central to our approach processes images of the model, garments, and accompanying text prompts, using garment images as the control factor. Internally, the network segregates into two streams for independent processing of model and garment data. These streams converge within a fusion network that facilitates the embedding of garment details onto the model's feature representation. On this foundation, we have established Outfit Anyone, comprising two key elements: the Zero-shot Try-on Network for initial try-on imagery, and the Post-hoc Refiner for detailed enhancement of clothing and skin texture in the output images.





Use Cases and Applications

Usage Scenario	Description
Social Media Platforms	Enable users to try on virtual makeup before purchasing, enhancing engagement and satisfaction.
E-commerce	Provide personalized makeup recommendations based on user preferences and skin tone, improving the shopping experience.
Entertainment Industry	Create realistic character transformations for movies, TV shows, and video games, reducing the need for physical makeup application.
Virtual Avatars	Generate lifelike avatars with customizable makeup styles for virtual worlds and metaverse environments.

Virtual Try-on represents a significant advancement in the field of makeup transfer, offering high-quality, diverse, and realistic results using diffusion models. By leveraging self-supervised learning, hierarchical texture decomposition, and iterative dual alignment, Virtual Try-on overcomes the limitations of traditional methods and opens new possibilities for creative expression in makeup design. As part of Alibaba Cloud's commitment to innovation, Virtual Try-on empowers businesses and individuals to explore new frontiers in digital beauty and personalization.

2.9 GENAI APPS - DIGITAL HUMAN

The Digital Avatar solution delivers superior interactivity and real-time response to enhance user journeys. High-quality avatars can be generated in batches for different scenarios and platforms, with compatibility across multiple devices and operating systems, web applications, and mini-programs. Leveraging Alibaba Cloud's AI capabilities including Qwen models, the avatars support intelligent dialogue and continuously improve comprehension, decision-making, and content generation capabilities.

2.9.1 ALIBABA CLOUD'S AVATAR

Alibaba Cloud Intelligent production enables the creation and training of digital avatars based on real humans, allowing for text-driven or speech-driven simulations of real-person broadcasting. Our avatar service allows you to create lifelike digital avatars that can be used for a variety of applications. These avatars are trained based on real human images and can simulate real-person broadcasting through text or speech inputs.

Key Features:

1. **Customizable Avatars:** Train avatars using real human images to create personalized digital representations.
2. **Text and Speech-Driven:** Generate videos by simply inputting text or speech, making content creation fast and efficient.
3. **High-Quality Output:** Professional Edition ensures high-quality avatars suitable for TV broadcasting and media, while the General Edition offers quick and cost-effective solutions.

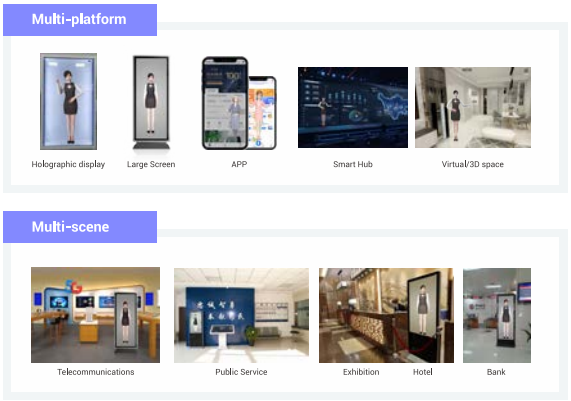
Why Choose Our Avatar Service?

- **Efficiency:** Quickly generate high-quality videos without the need for extensive recording sessions.
- **Flexibility:** Adapt avatars to various industries and use cases, from entertainment to corporate communications.
- **Cost-Effective:** Opt for the edition that best fits your budget and requirements, ensuring maximum value for your investment.

2.9.2 ANT DIGITAL AVATAR

The Ant Digital Avatar platform is an AI service for generating high-quality avatar videos and livestreams. It features conversational avatars that can be connected to a default knowledge base, custom dialog systems, or large language models like Qwen. The platform also offers solutions to clone your appearance and natural speaking voice, creating a digital copy of yourself. Additionally, you can integrate the platform's avatar capabilities into your product and programmatically create avatar videos using APIs.

Digital Humans: Gaining Market Recognition and Adoption



Key Features

- **Content Generation:** just provide a text/voice script, and then generate avatar videos in system
- **Avatar Livestreams:** support interactive live-streaming with digital avatar provide high efficiency, low cost, and uninterrupted service.
- **Instant Avatar service:** clone your appearance and natural speaking voice for customized avatar.
- **Conversational digital avatar:** connect the default Knowledge Base, your own dialog system or an LLM like Qwen.
- **API:** integrate platform's avatar capabilities into your product and then create avatar videos programmatically with API.

Advantages

- **High performance rendering mode:** the Ant Digital Avatar Platform is equipped with both client-side and cloud-side rendering mode. The client-side rendering technology is of high performance, high reliability and high compatibility.
- **Intelligent interactions:** the Ant Digital Avatars is built on our smart dialogue bot technology. it can also directly interface with LLM, continuously expanding the Digital Avatar's comprehension, decision-making, and content generation abilities
- **Easy to integrate:** integrates Avatar into online business processing by integrating client-side SDK.
- **Cross platforms:** can easily adapt to platforms such as Android, iOS, Web, WeChat, and Alipay Mini Programs. It also works well on IoT devices whith Android or Windows operation system, for instance, all-in-one machines, smartphones, PCs, and digital screens.

2.9.3 APPLICATIONS AND USE CASES

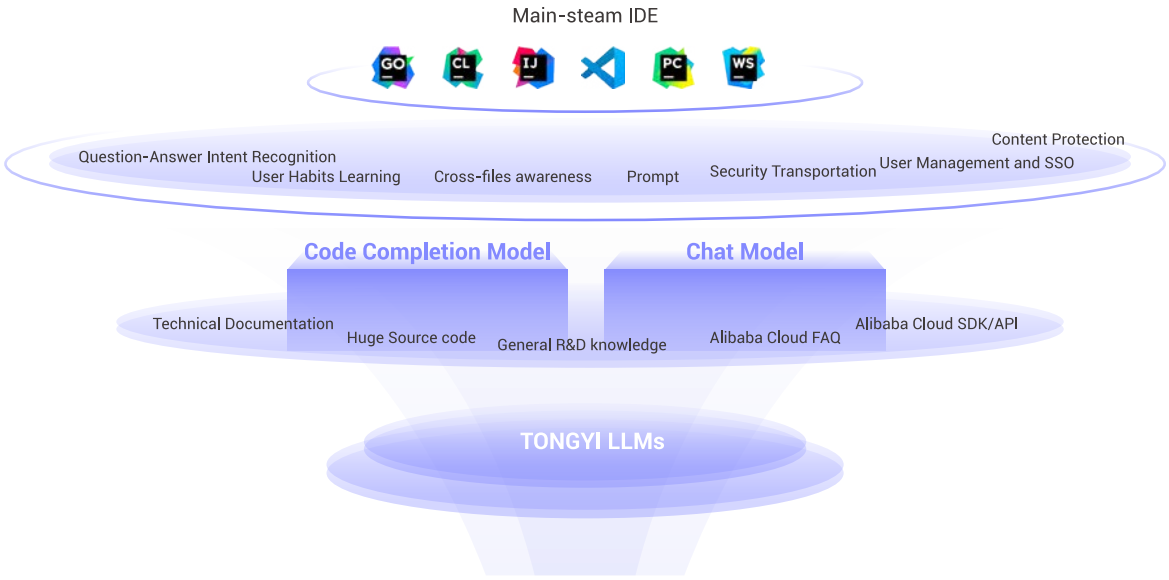
Use Case	Description
Avatar Content Generation	Generate high-quality avatar videos programmatically using APIs, ideal for creating engaging content for marketing, education, and entertainment.
Conversational Digital Avatars	Create interactive avatars that can engage in conversations by connecting them to a default knowledge base, custom dialog systems, or large language models like Qwen. Suitable for customer service, virtual assistants, and interactive storytelling.
Live-streaming Avatars	Utilize AI-driven avatars for live-streaming events, providing real-time interaction with audiences. Perfect for virtual events, product launches, and live broadcasts.
Digital Copy of Your Appearance and Voice	Clone your appearance and natural speaking voice to create a digital copy of yourself. This feature is useful for personal branding, virtual presence in meetings, and creating personalized content.

2.10 GENAI APPS - CODE ASSISTANT

2.10.1 TONGYI LINGMA – ALIBABA CLOUD INTELLIGENCE'S ADVANCED GENERATIVE AI

Tongyi Lingma is an intelligent programming assistant tool based on the Qwen large language models. It provides capabilities such as real-time continuation at the line or function level, natural language to code generation, unit test generation, code optimization, comment generation, code explanation, intelligent Q&A, and exception error troubleshooting. It is optimized for Alibaba Cloud's cloud service usage scenarios, helping developers code more efficiently and smoothly.

Overall Product Architecture



Supported Programming Languages

- Java, Python, Go, C#, C/C++, JavaScript, TypeScript, PHP, Ruby, Rust, Scala, and Kotlin.

Supported IDEs

- JetBrains IDEs (IntelliJ IDEA, PyCharm, GoLand, WebStorm, Android Studio, etc.)
- Visual Studio Code
- Visual Studio

Key Features of Tongyi Lingma

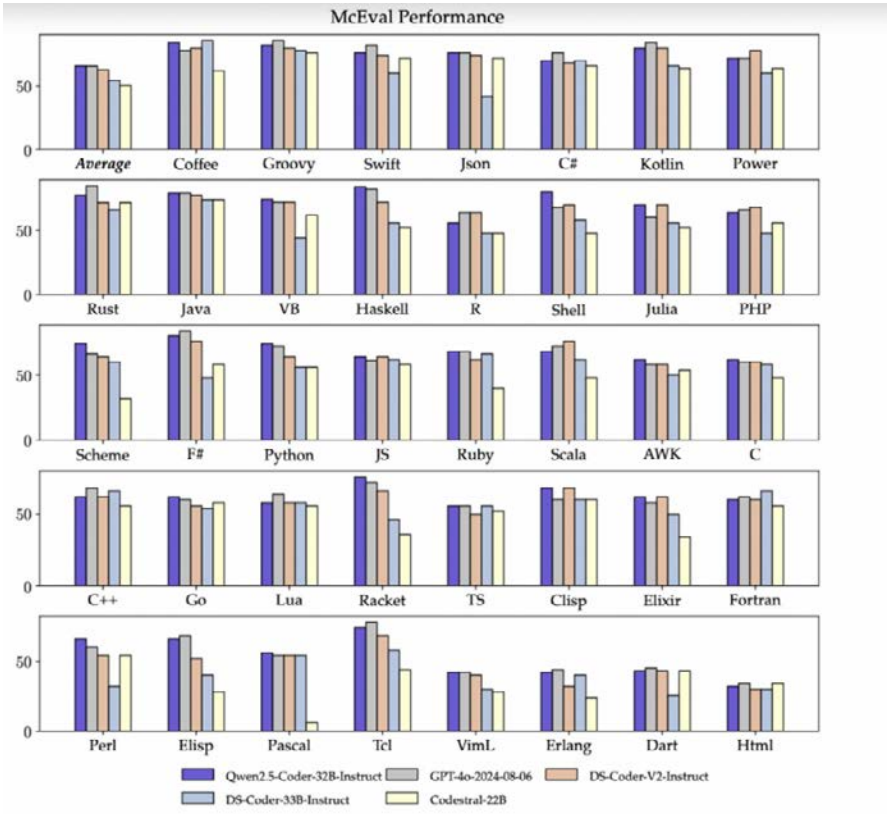
Industry	Description
In-line Code Generation	Real-time Line/Function Suggestions: Generates line and function suggestions based on current syntax and context.
	Code from Comments: Generate code directly from comment descriptions in the editor.
Intelligent Q&A for R&D	Coding Q&A: Get quick answers and solutions without leaving the IDE.
	Local Project Q&A: Understand and query code within the current project using natural language.
	Enterprise Knowledge Base Q&A: Use enterprise data to build a knowledge assistant for better collaboration.
Intelligent Generation	Unit Test Generation: Generate unit tests for various frameworks (JUnit, Mockito, etc.).
	Code Comment Generation: Automatically generate method and inline comments.
	Code Explanation: Identify and explain selected code in various programming languages.
	Commit Message Generation: Generate Git commit messages with one click.

Industry	Description
Coding Problem Resolution	Code Optimization: Identify and suggest fixes for coding issues and performance bottlenecks.
	Code Issue Fix: Provide one-click repair suggestions for code problems.
	Exception Error Troubleshooting: Quickly diagnose and suggest fixes for runtime exceptions.
Enterprise Management	Unified Authorization: Manage developer permissions centrally.
	Statistical Reports: Track developer activity and tool usage.
	Knowledge Management: Efficiently manage enterprise-specific data and knowledge.
	Dedicated VPC Deployment: Ensure secure and compliant deployment within enterprise networks.

Tongyi Lingma offers developers an intelligent development experience and leads the new paradigm of AI-native development.

2.10.2 QWEN2.5 - CODER - THE SOTA OPEN-SOURCE CODE MODEL

The Qwen2.5-Coder series represents a pivotal milestone in Alibaba Cloud's commitment to advancing the field of generative AI. This innovative lineup of models is specifically designed for code-related tasks, boasting unparalleled capabilities in code generation, repair, reasoning, and multi-language support.



source: https://qwenlm.github.io/blog/qwen2.5-coder-family/

Key Capabilities

1. **Powerful Coding Proficiency:**

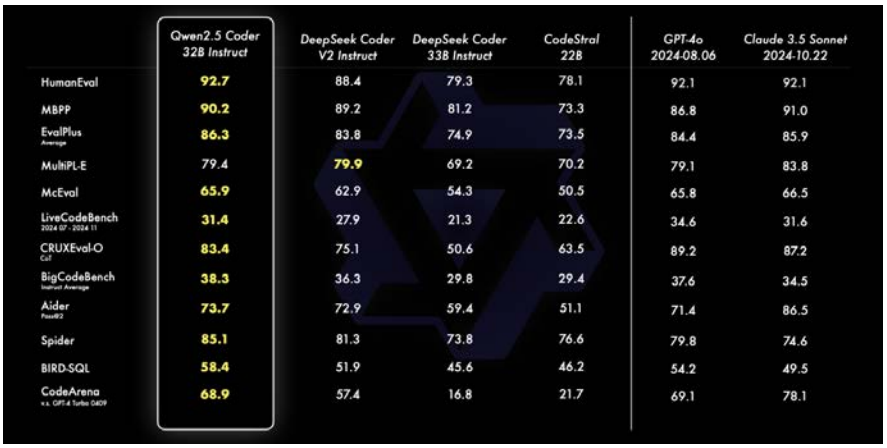
- **Model:** Qwen2.5-Coder-32B-Instruct, SOTA performance
- **Highlights:** Strong in code generation, repair (73.7 Aider), and reasoning

2. **Multi-Language Expertise:**

- **Languages:** Over 40, including niche languages like Haskell and Racket
- **Scores:** 65.9 McEval (understanding), 75.2 MdEval (repair)

3. **Scaling Excellence:**

- **Sizes:** 0.5B, 1.5B, 3B, 7B, 14B, and 32B parameters
- **Advantages:** Optimized for developers and enterprises, balancing complexity and efficiency



	Qwen2.5 Coder 32B Instruct	DeepSeek Coder V2 Instruct	DeepSeek Coder 33B Instruct	CodeStral 22B	GPT-4o 2024-08.06	Claude 3.5 Sonnet 2024-10.22
HumanEval	92.7	88.4	79.3	78.1	92.1	92.1
MBPP	90.2	89.2	81.2	73.3	86.8	91.0
EvalPlus	86.3	83.8	74.9	73.5	84.4	85.9
MultiPL-E	79.4	79.9	69.2	70.2	79.1	83.8
McEval	65.9	62.9	54.3	50.5	65.8	66.5
LiveCodeBench	31.4	27.9	21.3	22.6	34.6	31.6
CRUXEval-O	83.4	75.1	50.6	63.5	89.2	87.2
BigCodeBench	38.3	36.3	29.8	29.4	37.6	34.5
Aider	73.7	72.9	59.4	51.1	71.4	86.5
Spider	85.1	81.3	73.8	76.6	79.8	74.6
BIRD-SQL	58.4	51.9	45.6	46.2	54.2	49.5
CodeArena	68.9	57.4	16.8	21.7	69.1	78.1

2.10.3 USE CASES AND APPLICATIONS

Generative AI coding assistants like Tongyi Lingma are widely adopted. A Q3 2023 Gartner survey of 598 global companies showed that 63% are piloting, deploying, or have deployed AI code assistants. Gartner's April 12, 2024, report noted that in early 2023, less than 10% of enterprise software engineers used these tools, but this is expected to reach 75% by 2028, making them a key asset for R&D efficiency.[8]

Industry	Use Case	Description
Software Development	Code Generation & Debugging	Automates routine coding tasks, identifies and fixes bugs, and enhances productivity.
Finance	Risk Management & Compliance	Assists in developing algorithms for risk assessment, compliance checks, and fraud detection.
Healthcare	Medical Software Development	Helps in building secure, compliant healthcare applications, including patient management systems.
Retail	E-commerce Platform Development	Facilitates the creation of robust e-commerce platforms with features like inventory management and customer analytics.
Manufacturing	Industrial Automation	Supports the development of automation scripts and software for managing production lines and quality control.
Education	Educational Software & Platforms	Aids in creating interactive learning platforms and educational tools for students and educators.
Telecommunications	Network Management & Optimization	Assists in Assists in developing network monitoring and optimization tools to ensure efficient communication infrastructure. communication infrastructure. infrastructure.
Energy	Smart Grids & IoT Applications	Helps in building applications for smart grid management, energy consumption monitoring, and IoT device integration.
Automotive	Autonomous Vehicle Software	Supports the development of software for autonomous driving systems, including navigation and safety features.
Entertainment	Game Development & Content Creation	Enhances game development processes and assists in creating interactive content for streaming platforms.

3. SECURITY AND PRIVACY ARE PRIORITIES

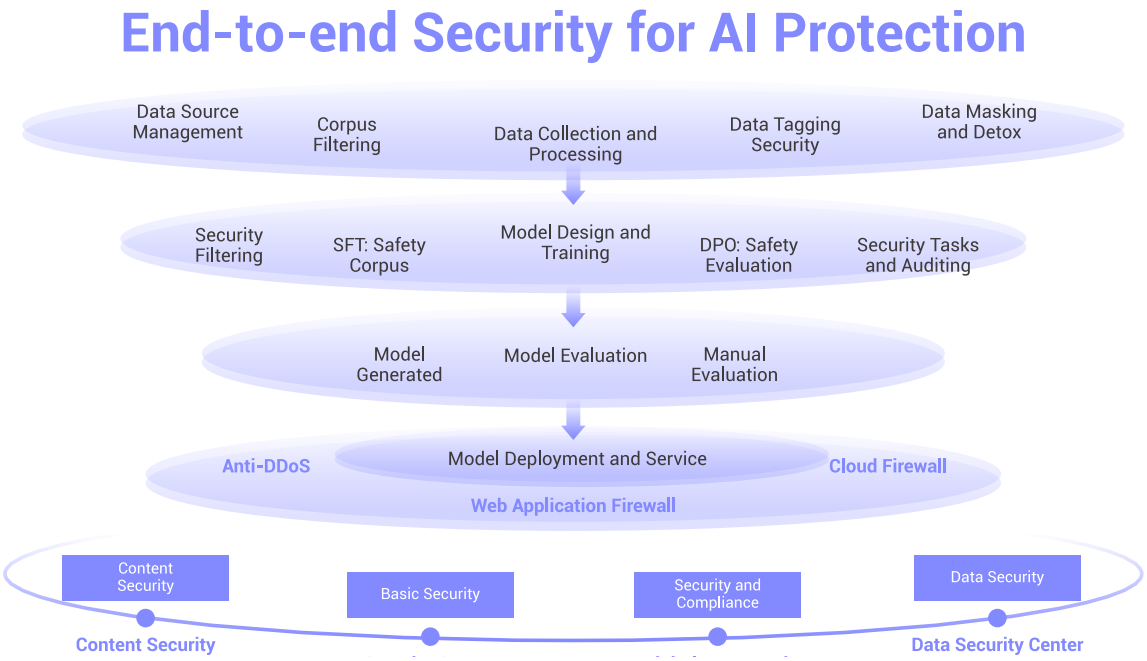
In the rapidly evolving landscape of Generative AI, ensuring the security and privacy of user data stands as a paramount concern. As models grow more sophisticated and are deployed across various industries, it is crucial that these systems operate within frameworks that prioritize the confidentiality, integrity, and availability of data. This chapter delves into the measures taken to safeguard data in cloud-based Generative AI services, focusing on Alibaba Cloud's approach to addressing key challenges.

Data Security Challenges

Generative AI models, especially those hosted on cloud platforms, present unique data security challenges:

1. **Security of Data on the Cloud:** With user data moving alongside models to the cloud, the platform must maintain data security levels equivalent to those of the user's private domain. Ensuring data is not exposed or leaked through model outputs requires stringent safeguards during transmission, storage, and application processes.
2. **Controllability of Model Data Usage:** Users need transparency and control over how their data is used by models. Unauthorized use of data for training can compromise confidentiality and diminish its commercial value, leading to user caution towards model applications.

3. **Auditable Operations and Traceable Responsibility:** When processing user data, there must be clear agreements on responsibilities and obligations between parties. The ability to audit actions and trace responsibility is critical for accountability, especially when identifying the root cause of issues and security weaknesses.



Countermeasures Implemented by Alibaba Cloud

To mitigate these challenges, Alibaba Cloud has implemented comprehensive countermeasures:

- **Reliable Platform:** Leveraging Alibaba Cloud's robust cloud computing infrastructure, which ensures high availability, stability, and rapid response to potential threats. The platform also supports compliance with various international standards, thereby providing a secure foundation for LLM services.
- **Trusted Link:** Utilizing advanced encryption techniques for data transmission and storage, along with secure protocols for API calls. Alibaba Cloud employs rigorous access controls and authentication mechanisms to prevent unauthorized access.
- **Controllable Data:** Offering users granular control over their data, including informed consent for processing and options for data deletion. Alibaba Cloud's models are designed to minimize the risk of reverse inference from outputs, protecting sensitive information.

- **Autonomous Options:** Providing flexible configuration options for data handling, such as masking sensitive data and isolating compute environments. Users can customize security settings based on their specific needs, enhancing both functionality and protection.
- **Auditable Operations and Traceable Responsibility:** Implementing detailed logging and monitoring systems to track all operations involving user data. This allows for transparent management and verification, ensuring full auditability of processes and multi-party rights.

Alibaba Cloud's commitment to data security extends beyond technical measures. By undergoing rigorous third-party audits and continuously updating its security policies, Alibaba Cloud aims to build trust among users and stakeholders alike. The company's dedication to maintaining a secure and compliant environment underscores its role as a leader in the Generative AI space, setting benchmarks for others to follow.

In addition to this document, Alibaba Cloud is pleased to offer two comprehensive security white papers designed to provide deeper insights into our commitment to security and data privacy. The first one, Alibaba Cloud Security Whitepaper 2024 , offers an in-depth overview of our latest security strategies, technologies, and best practices implemented across our cloud infrastructure. This white paper is essential reading for anyone interested in understanding how we protect our customers' data and ensure the highest levels of security.

The second white paper, Alibaba Cloud LLM Service Data Privacy and Security White Paper , focuses specifically on the security measures and data privacy protocols associated with our Large Language Model (LLM) services. It details the robust frameworks we have established to safeguard user data while leveraging advanced AI capabilities, ensuring both innovation and protection.

4. CUSTOMER STORIES

AstraZeneca with OpenTrek Platform

> We ultimately opted for Alibaba Cloud because their solution for adverse effects reporting outperformed competitors.
>
> **-- Xin Zhong,**
> **IT Head of AstraZeneca China**

A global biopharmaceutical company focused on science-led innovation, AstraZeneca aims to transform healthcare by leveraging scientific advancements and emerging technologies. The company integrates data science and AI into its R&D to enhance drug development and understanding of diseases. In 2024, AstraZeneca established a new strategic R&D center in Shanghai, marking a significant step in its global expansion.

Challenges

AstraZeneca has expanded its global presence by establishing R&D centers and manufacturing facilities in China to utilize local capabilities and market opportunities. A key imperative for multinational pharmaceutical companies in China is complying with local regulatory requirements, such as reporting adverse drug reactions (ADR) to the National Medical Products Administration (NMPA). This process is time-consuming for medical researchers, who must gather and analyze medical literature from healthcare professionals and institutions to create reports that meet strict regulatory standards, including specific formatting and editorial guidelines.

Why Alibaba Cloud

Alibaba Group is a key player in digitalization in China and globally, with a strong history of integrating AI across multiple industry sectors. Alibaba Cloud, the technology backbone of the group, plays a significant role in AI-driven innovation in China, supporting over 80% of China's tech businesses and more than 50% of LLM-based innovations. A leading LLM in China and globally, Tongyi Qwen offers various model sizes and multi-modal features, and is ideal for cross-lingual scenarios.

Architecture

Working with Alibaba Cloud, AstraZeneca's team has built the industry's first adverse event summary system based on Tongyi Qwen LLM and Opentrek Platform. The system makes full use of the AI engineering capabilities provided by Alibaba Cloud's Opentrek Platform, integrating AstraZeneca's private knowledge bases and publicly available medical and pharmacological knowledge.

With Alibaba Cloud's Opentrek Platform, both development and administration tasks are now more streamlined.

AstraZeneca's researchers can create tasks and input documents into the system, which generates an adverse event report, based on the source information. The report only needs a human review and minimal edits before sign-off.

Private deployment on the cloud, leveraging Opentrek Platform, allows AstraZeneca to leverage the scalability and stability of using the cloud-based solution. This ensures a more stringent data and system segregation, while crucially safeguarding data privacy.

Key Results

Thanks to Alibaba Cloud's Opentrek Platform, accuracy of the adverse event reports has increased from 90% to 95%.

As the system can quickly comb through vast amounts of medical literature and formulate output documents according to the requirements, efficiency in creating the adverse event reports has increased by 300%.

Looking Forward

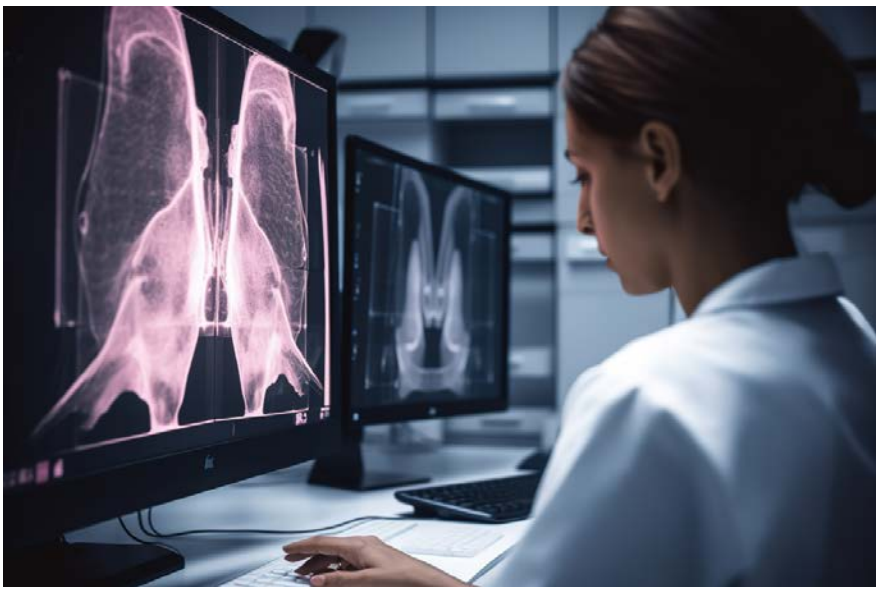
Working with Alibaba Cloud, AstraZeneca has proudly made this pioneering solution the first of its kind in the pharmaceutical industry, which will empower pharmaceutical businesses and enhance safety for patients. The company is looking forward to pioneering more AI-driven innovations, by leveraging its industrial expertise in concert with Alibaba Cloud's cutting-edge technologies for greater digitalization.

Haleon with Tongyi Qwen

HALEON

> We are honored to collaborate with Alibaba Cloud on this venture.
>
> **-- Susan Gu, General Manager, Haleon Mainland China and Hong Kong**

Haleon (LSE / NYSE: HLN) is a global leader in consumer health, with a purpose to deliver better everyday health with humanity. Haleon's product portfolio spans five major categories - Oral Health, Pain Relief, Respiratory Health, Digestive Health and Other, and Vitamins, Minerals and Supplements (VMS). Its long-standing brands - such as Advil, Sensodyne, Panadol, Voltaren, Theraflu, Otrivin, Polident, parodontax and Centrum - are built on trusted science, innovation and deep human understanding.



Challenges

The healthcare landscape is evolving, with a growing emphasis on personalized consumer experiences. Haleon, built on the foundation of scientific expertise, innovation, and deep human understanding, aims to harmonize consumer insights with cutting-edge healthcare service innovation. To ensure a premium customer experience for its vast consumer base, the company sought to implement a fully automated, real-time service delivery system. This technologically-rich solution would address each consumer's enquiry and demand instantly, eliminating delays and enhancing overall satisfaction.

Haleon aimed to develop an intelligent, responsible, and professional system that could handle a large volume of consumer enquiries in real time, ensuring a premium customer experience without compromising personalization or efficiency.

Why Alibaba Cloud

With healthcare at the forefront of the AI revolution, the nutrition, medicine development, and health management industries have largely begun leveraging Large Language Models (LLMs) to unlock unprecedented value.

By harnessing Alibaba Cloud's sophisticated LLM, Tongyi Qianwen (Qwen), and integrating it with Haleon's proprietary knowledge graph through retrieval-augmented generation, Haleon's solution now seamlessly integrates across all customer touchpoints, ensuring a cohesive end-to-end journey and hyper-personalized healthcare management.

Architecture

Leveraging the Qwen large language model, integrated with Haleon's existing knowledge graph, Alibaba Cloud supported Haleon to build AI nutrition assistant services through retrieval-augmented generation. The solution includes end-to-end design, development, and integration across the stakeholder spectrum and implements the following components:

- **Qwen Large Language Model and Prompt Engineering:** Through Prompt Engineering, the input and output of the Qwen model are arranged and connected to realize question pre-processing, personalized Q&A based on user tags, structured nutritionist responses, specified tone styles, and other features.
- **Vector Recall:** Sort out and split the data from the enterprise's exclusive knowledge base, and vectorize the knowledge through the Embedding model. The purpose is to recall unstructured knowledge based on user questions and use it to enhance the answers of the large language model.
- **Knowledge Graph Recall:** Knowledge graph recall makes full use of the existing data assets of the enterprise and uses NL2Cypher to retrieve knowledge such as relationships and attributes that are beneficial to question answering from the knowledge graph. This improves the professionalism and reliability of the large language model's answers.
- **Recall Sorting:** Integrate multi-channel recall results (including vector recall, sub-graph recall) to ensure the reliability of the retrieval results while eliminating any noise, thereby enhancing the answering effect of the large base model.
- **Front-end Integration:** By integrating Qwen's standard API interface with the front-end system, multi-touch service integration is achieved, particularly with Enterprise WeChat. This integration provides functions including login, authorization and authentication, messaging, and security encryption, among others.

Key Results

The partnership between Alibaba Cloud and Haleon, powered by Tongyi Qwen, has revolutionized personalized healthcare for millions. This collaboration enabled Haleon to achieve unprecedented precision and insight in addressing health inquiries

Nutritionists utilizing the AI nutritional assistant experienced a substantial enhancement in professionalism and scientific rigor, alongside notable gains in service efficiency. Currently, a single

nutritionist supported by this technology can serve over 1,300 consumers, representing a six-fold increase in efficiency compared to traditional models where human nutritionists typically handle around 200 clients each on other platforms. Additionally, they were able to respond to inquiries in the user's preferred language.

Looking Forward

Haleon plans to further enhance its AI-powered health management services by expanding its capabilities of knowledge graph x LLM and improving personalization. Leveraging Alibaba Cloud's scalable computing resources and proprietary AI frameworks, Haleon aims to accelerate innovation in consumer healthcare, emphasizing data-driven insights and intelligent automation throughout its value chain.

In addition to the AI-powered nutritional assistant project already underway, Haleon also looks forward to collaborations involving AI-generated content, AI brand ambassadors, AI-personalized experiences, and exploring how to apply large language model technologies in conjunction with Haleon's trusted science to empower healthcare professionals.

Shiseido with Tongyi Qwen

SHISEIDO

DRUNK ELEPHANT™

> Selecting Alibaba Cloud Services as our partner was a wise decision rooted in their demonstrated expertise in data analysis, AI services, and language models. Their understanding of applications and services has been pivotal in enhancing our consumer experience and technology innovation. Moreover, Alibaba's extensive experience in driving growth within the retail and digital sectors, coupled with their technological capabilities, aligns perfectly with Drunk Elephant's vision for creating innovative experiences for their consumers.
>
> **-- Tina Chen, Chief Digital Officer, Shiseido China**

In 2019, Shiseido acquired Drunk Elephant, one of the fastest growing skincare brands in the last decade. Drunk Elephant is a highly effective skincare brand with research proven results, established in the United States in 2013.

Founded by Tiffany Masterson, Drunk Elephant is renowned for its ingredient-elimination philosophy. Drunk Elephant newly launched in Chinese Mainland this April, rolling out in 270 Sephora stores across the country, where consumers can experience the brand's award-winning products such as Protini Polypeptide Cream, Virgin Marula Luxury Facial Oil, and B-Hydra Intensive Hydration Serum here. Drunk Elephant is beloved by dermatologists, beauty editors, celebrities, and KOLs around the world, and has won over 450 awards globally.

Challenges

Drunk Elephant has an engaged global community that the brand has built for over ten years. Drunk Elephant has a strong connection with its community and encourages consumers to "Listen to your skin" in order to find the right Drunk Elephant products based on their skin needs. Drunk GPT invites the Chinese consumer to be part of the Drunk Elephant community and learn about the brand's unique skin philosophy and discover easy, customized skincare routines.

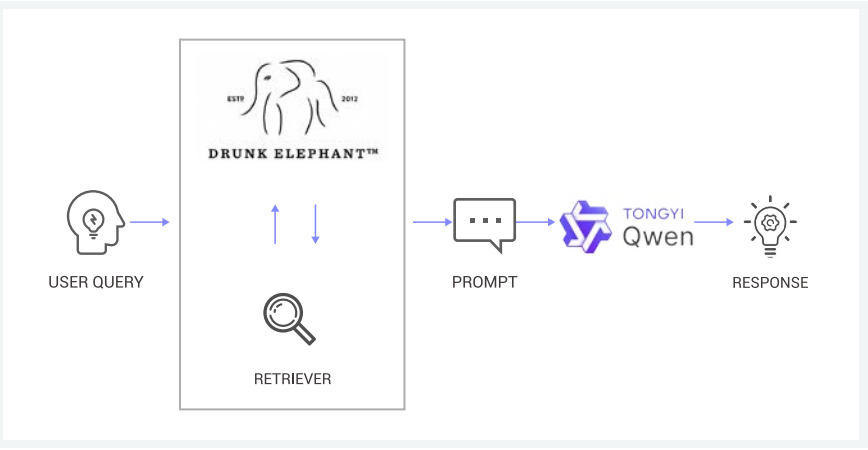
Hence, when Drunk Elephant launched in Chinese mainland, it looked to leverage cutting-edge technology to provide a highly interactive and memorable customer experience in its Wechat Community and offline activities.

Why Alibaba Cloud

Drunk Elephant partnered with Alibaba Cloud because it needed to work with a company that had cutting-edge technology, professionalism, and good leverage in China. Alibaba Cloud was familiar with the client's industry and IT infrastructure, and, as the largest cloud computing company in China and Asia Pacific, offered digital expertise and a huge advantage in the Eastern market.

Alibaba Cloud explored the technical feasibility and creative ideas for AI-generated content (AIGC) engagement and provided consultation on retrieval-augmented generation (RAG) to ensure Drunk Elephant's needs were met. Alibaba Cloud offered project planning, including a proof-of-concept plan and a resource support plan, and worked with the client and delivery team on the acceptance criteria to ensure a quick and efficient workflow.

Architecture



To answer Drunk Elephant's needs, Shiseido worked with Alibaba Cloud to build "Drunk GPT," an interactive multi-round dialogue LLM that understands text, audio, and images and focuses on skincare knowledge, relevant Q&A, product consultation, and product recommendations. Drunk GPT operates via two branches: Directly on Drunk Elephant's WeChat mini program and through the GPT chat on the Drunk Elephant webpage.

Drunk GPT is deployed on an Alibaba Cloud Container Service for Kubernetes (ACK) cluster, its content response is powered by Alibaba Cloud's Tongyi Qwen LLM, and it uses NLS to provide text-to-speech conversion. Alibaba Cloud's Tongyi Qwen LLM, which is used by Drunk GPT, has surpassed the world's leading AI technologies in assessments for Chinese language tasks and achieved the highest score for open source models in MMLU evaluation for English language tasks.

This solution based on retrieval augmented generation (RAG) and multi-agent workflows can effectively integrate Drunk Elephant's private knowledge, such as product information, and trigger agentic workflows based on the user's interactions, such as requesting the user to participate in a questionnaire. Finetuning was performed on top of the Qwen LLM to incorporate industry-specific knowledge and to deliver higher accuracies when responding to consumers' queries.

Key Results

In addition to the Drunk GPT interactive chatbot, Drunk Elephant's partnership with Alibaba Cloud has resulted in improved online marketing and security.

With the help of Alibaba Cloud, Drunk Elephant runs interactive campaigns on the Drunk Elephant WeChat mini program, including attracting new customers through interactive AI picture generation challenges and running services for consumers via trained industrial models for brands. An example of the latter is allowing a space for AI to provide professional and interactive feedback to customers. These functionalities require multiple interactions with the app.

Alibaba Cloud has also built multiple security mechanisms that work with Shiseido's high security standards in IT infrastructure. These content security measures will prevent abuse, such as the generation of inappropriate content, whether requested or retrieved.

Looking Forward

Moving forward, the partnership between Shiseido and Alibaba Cloud for Drunk Elephant will focus on enhancing DrunkGPT, innovating input/output methods and information retrieval, and growing its capabilities in the Drunk Elephant WeChat mini program. Moreover, as AI and cloud technology improves, Shiseido will continue working with Alibaba Cloud to explore more AI-based innovations.

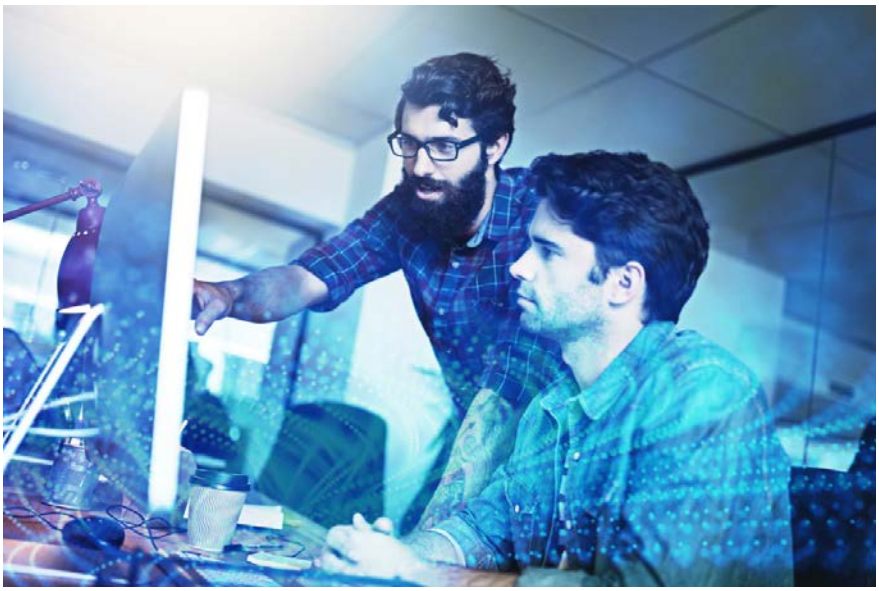
X-Verse Technologies with Tongyi Qwen

X-VERSE 源宇

> Alibaba Cloud has helped us enter international markets such as Greater China, ASEAN, and ANZ. Leveraging Alibaba Cloud's industry expertise and global reputation, we are able to meet both local and international ISO standards for data security and service SLAs.
>
> **-- Gerinna Wang,**
> **Founder and CEO of X-Verse Technologies**

The X-Verse design and creativity ecosystem is a collaborative, automated platform that connects creators and stakeholders in real-time, making creation more accessible and executable. As a pioneer in ASEAN, X-Verse is leading the way in utilizing AI agents and multimodal HCI (Human Computer Interaction) for design and creativity generation.

X-Verse has also invested in a cutting-edge technology-driven platform that offers a collaborative experience for creators and stakeholders alike. X-Verse incorporates innovative technologies, including large language models, AI-driven workflows, and multimodal human-computer interactions, to maintain its position as an industry pioneer.



Challenges

Operating in Asia with an international footprint covering Greater China, ASEAN, and ANZ, X-Verse faced the challenge of finding a reliable cloud service partner to ensure seamless connectivity across these diverse regions and globally. As a startup, X-Verse strongly emphasizes managing capital expenditure (CAPEX) wisely, seeking solutions that offer flexible cloud hosting services at an economical cost. Lastly, X-Verse uses large language models in its services, so it wanted a reliable, open-sourced LLM.

Why Alibaba Cloud

X-Verse selected Alibaba Cloud for several compelling reasons. Firstly, Alibaba Cloud's extensive international presence and strong reputation, particularly in China, provide a secure framework to meet local ISO standards, ensuring compliance with rigorous data security and service level agreements (SLAs).

Secondly, Alibaba Cloud's Qwen 1.5/2.0 offers one of the industry's leading open-source large language models (LLMs), with capabilities such as Chinese language operations, Retrieval-Augmented Generation (RAG), and text-to-image functionalities. Its transformative potential can help X-verse to revolutionize traditional search and content creation methods.

Lastly, Alibaba Cloud's flexible cloud hosting services provide X-Verse with an economical choice.

Architecture

X-Verse utilized various solutions from Alibaba Cloud, including Universal Type U1, a versatile instance type of Elastic Compute Service (ECS) offering scalable computing power for diverse workloads, and Object Storage Service (OSS), Alibaba Cloud's robust storage solution providing secure, scalable, and cost-effective object storage for enterprises and developers.

It also implemented Elastic Block Storage (EBS), a reliable block-level storage offering high-performance and low-latency access for applications running on Alibaba Cloud. Additionally, it used a Dynamic Content Delivery Network (DCDN), which accelerates content delivery with high efficiency and low latency across global networks, enhancing the user experience.

Furthermore, X-Verse leveraged Qwen 1.5-72B (open-source), a cutting-edge large language model (LLM) on Alibaba Cloud, trained on extensive multilingual data to empower advanced natural language processing and generation capabilities.

Key Results

Partnering with Alibaba Cloud enabled X-Verse to transform business operations in the following ways:

- **Transforming Knowledge Acquisition Experience:** X-Verse utilized Alibaba Cloud's AI technology, including large language models (LLMs) and AI, to transform the ideation ecosystem. This transformation streamlined knowledge gathering and the entire chain of content management from content creation and analysis to distribution, offering customers a comprehensive information repository covering text, images, and videos. X-Verse also leverages LLM to conduct real-time searches to acquire knowledge (image, text, video) and provide mind-mapping guidance to end users. Its mind-mapping agentic workflows facilitated better organization of ideas and guided actionable steps, significantly enhancing content quality and efficiency.

- **Improving Decision-Making Efficiency:** By integrating large language models and pre-built agentic workflows for information processing, X-Verse helped end users achieve significant improvements in decision-making processes. This innovation reduced the time and costs traditionally associated with multiple search and decision steps. X-Verse's agentic workflows guided customers seamlessly from information retrieval to execution and booking, optimizing design processes and making design more intelligent and efficient.
- **Fostering Innovative Thinking Approach:** X-Verse pushed the boundaries of imagination and creativity through AI to transform the fields of smart sales and marketing. By harnessing intelligent data analysis and machine learning, X-Verse tries to deepen business understanding of market dynamics and consumer demands, shaping robust sales and marketing strategies accordingly.

Looking Forward

The rapid evolution of AI is driving digital transformation across all industries. The potential for generative AI adoption in the design and creativity ecosystem is immense. Digital intelligence is revolutionizing how we interact with the world, making us more productive and enabling smarter decision-making.

X-Verse's collaborative design and creativity ecosystem, combined with Alibaba Cloud's infrastructure and large language model solutions, introduces groundbreaking methods for knowledge acquisition, execution, and seamless connectivity among all stakeholders in the design and creativity ecosystem.

Looking ahead, this partnership will further drive digital transformation, unlocking vast opportunities for generative AI adoption within the design and creativity ecosystem.

Lightblue with Tongyi Qwen

> The advent of generative AI has enabled software to perform tasks that were previously only performed by human beings.
>
> **-- Atomu Sonoda, Founder, Lightblue**

Lightblue Co., Ltd. is a startup dedicated to democratizing AI. It has launched LLab, a specialized team focused on the research and development of generative AI and large language models (LLMs), prioritizing safety and transparency. Lightblue's mission is to broaden the applications of AI technology and drive transformative, positive change in society.

Challenges

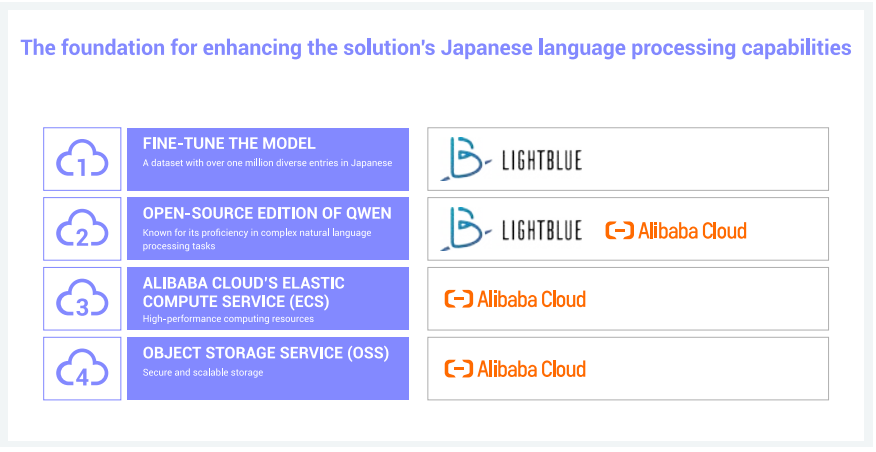
Following the release of ChatGPT, the popularity of generative AI surged. Despite the growth, no domestically produced LLMs existed in Japan, leading Lightblue to initiate its own LLM development.

Lightblue introduced its LLMs, Karasu and Qarasu, which stood out due to the rigorous efforts of engineers who refine and optimize data sets daily, ensuring exceptional performance in Japanese. However, the major challenge was finding the most suitable base LLM specifically for training in the Japanese language. Lightblue wanted a technology partner capable of addressing this challenge and providing the necessary expertise and support.

Why Alibaba Cloud

The support of Alibaba Cloud's Tongyi Qianwen (Qwen) to Lightblue was crucial for the release of its Karasu and Qarasu models around December 2023. Lightblue explored prominent models like LLama2 and Mistral Large to assess the best foundation for development. While both showed strong performance in English, Qwen emerged as the best choice for handling Japanese. Its advanced architecture and extensive training in East Asian languages provided outstanding accuracy when fine-tuned for Japanese, ensuring clear and relevant interactions. Alibaba Cloud's Qwen was crucial in giving Lightblue the capabilities needed to succeed in Japanese language processing.

Architecture



Lightblue primarily utilized the open-source edition of Tongyi Qianwen (Qwen), Alibaba Cloud's advanced foundation LLM, known for its proficiency in complex natural language processing tasks and strong multilingual capabilities.

In addition, Lightblue employed several other Alibaba Cloud solutions. The Elastic Compute Service (ECS) was used to offer scalable, on-demand computing resources for rapid deployment of virtual servers, featuring various instance types and storage options for optimal performance. Alibaba Cloud's Server Load Balancer (SLB) was also used to distribute incoming traffic across multiple servers, ensuring the high availability and reliability of the application.

Lightblue also utilized Alibaba Cloud Object Storage Service (OSS), which provides scalable and secure storage for large amounts of unstructured data, ensuring reliable data management.

Key Results

Qwen offers a range of model sizes, from lightweight to large, which proved highly convenient during development. While a 72b model is more accurate for achieving high scores, a 7b model is easier to manage and comes with various parameter sizes, including output speed, making it user-friendly from a development standpoint.

As Qwen was released on Hugging Face, it was in a state that allowed for seamless learning, enabling Lightblue to develop its LLM without any obstacles.

Looking Forward

Lightblue and Alibaba Cloud anticipate collaborating on more projects in the near future. Lightblue aims to enhance the accuracy of its SaaS service, Lightblue Assistant, a RAG-based solution, to eventually replace the parts that currently rely on APIs provided by other vendors. Additionally, Lightblue is considering developing a user-friendly LLM model tailored for customers who need to run it in a local environment.

5. REFERENCES

1. **McKinsey&Company**

The economic potential of generative AI: The next productivity frontier

https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier

2. **Stuart J. Russell and Peter Norvig**

Artificial Intelligence: A Modern Approach. Pearson

https://www.pearson.com/en-us/subject-catalog/p/artificial-intelligence-a-modern-approach/P200000003500/9780137505135

3. **Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2023)**

Retrieval-Augmented Generation for Large Language Models: A Survey

https://arxiv.org/abs/2312.10997

4. **Alibaba's Tongyi Lab**

Generalizing an LLM from 8k to 1M Context using Qwen-Agent

https://qwenlm.github.io/blog/qwen-agent-2405/

5. **Alibaba's Tongyi Lab**

Qwen2-VL: To See the World More Clearly

https://qwenlm.github.io/blog/qwen2-vl/

6. **Alibaba's Tongyi Lab**

QVQ: To See the World with Wisdom

https://qwenlm.github.io/blog/qvq-72b-preview/

7. **Alibaba's Tongyi Lab**

QwQ: Reflect Deeply on the Boundaries of the Unknown

https://qwenlm.github.io/blog/qwq-32b-preview/

8. **Gartner**

Gartner Says 75% of Enterprise Software Engineers Will Use AI Code Assistants by 2028

https://www.gartner.com/en/newsroom/press-releases/2024-04-11-gartner-says-75-percent-of-enterprise-software-engineers-will-use-ai-code-assistants-by-2028

ABOUT

Established in 2009, Alibaba Cloud (alibabacloud.com), the digital technology and intelligence backbone of Alibaba Group, is among the world's top three IaaS providers, according to Gartner. It is also the largest provider of public cloud services in China, according to IDC.

Alibaba Cloud provides a comprehensive suite of cloud computing services to businesses worldwide, including merchants doing business on Alibaba Group marketplaces, start-ups, corporations and public services.

Alibaba Cloud is the official Cloud Services Partner of the International Olympic Committee.

www.alibabacloud.com/contact-sales