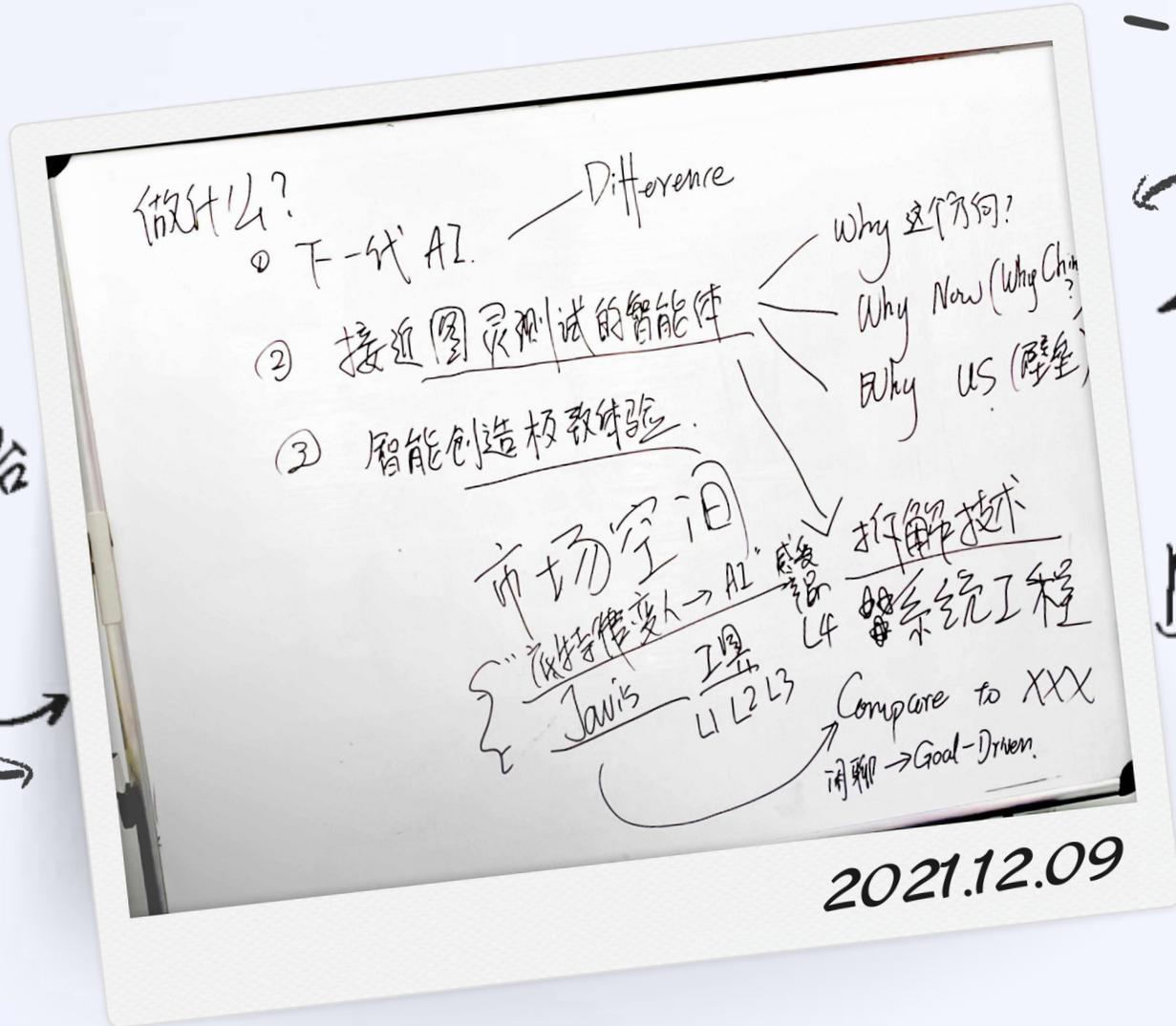


MiniMax

Intelligence with Everyone

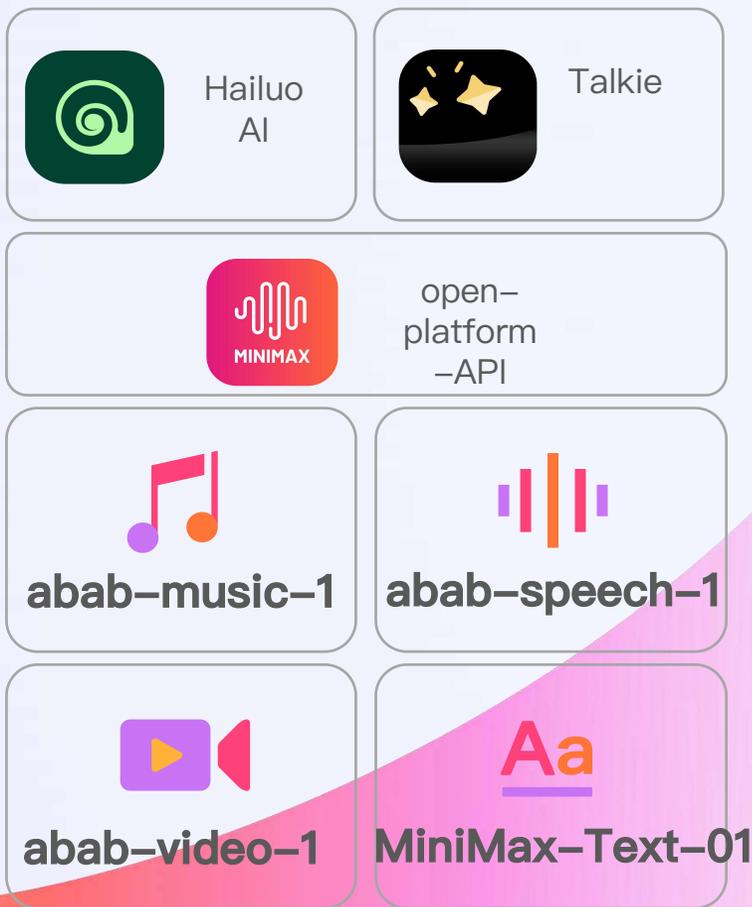
1,000 Days of Entrepreneurship without Shortcuts

回到最开始的地方



MiniMax
Day One

From 0 to 5 Billion Interactions per Day



MiniMax Daily

7 Trillion
Tokens

7,000 lifetimes
in 1 day

25 Million
Images

500x
Forbidden City's
paintings

600,000
Hrs Audio

60,000 books
listened in 1 day

Millions
Video Clips

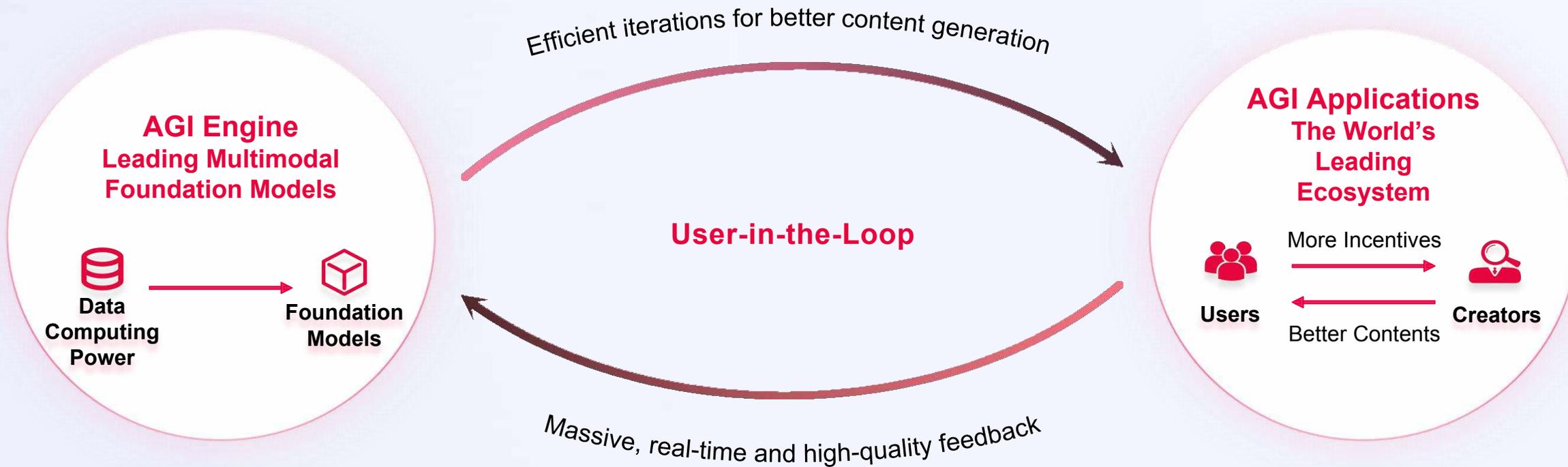
World-leading
scale

Day 1

Day 996

Sources: Company data from MiniMax; Official collection of 50,000 paintings from the Forbidden City; Assuming each book takes 10 hours to read aloud; Assuming one person has 1 billion tokens of interaction in a lifetime

A Highly-effective Flying Wheel between Foundational Model and User Feedback Building the “Super APP Factory” in the AGI Era



Text



Video



Voice



Music



Xingye



Talkie



Hailuo AI



Open Platform

MiniMax – LLM Leader in APAC: ToC + ToB Dual-Engine with a Global Footprint



Technology



- Daily query volume **ranked top 2 globally (second only to OpenAI), processing 5B+ queries per day**
- Processing **7 trillion** tokens daily, **50%** of the scale of **OpenAI**
- **1st in APAC**; global leading full-stack multimodal models that can be commercially used besides OpenAI and Google, including: GPT4-level text-based LLM, multimodal model similar to GPT-4o using **MoE + Linear Attention** technology, production-level video model, hyper-realistic voice model and music model

Talent



- Core large model R&D team composed of members from **the world's leading AI institutions**, including Meta, Microsoft, Llama, Mistral, Google, etc., and **over 40%** of them have overseas backgrounds
- The R&D leader possesses expertise in managing a development team comprising more than **a thousand members**; the ToC team has overseen products achieving **150 million daily active users**; and the ToB team has guided products that have generated revenues **exceeding 2 billion dollars**

Business

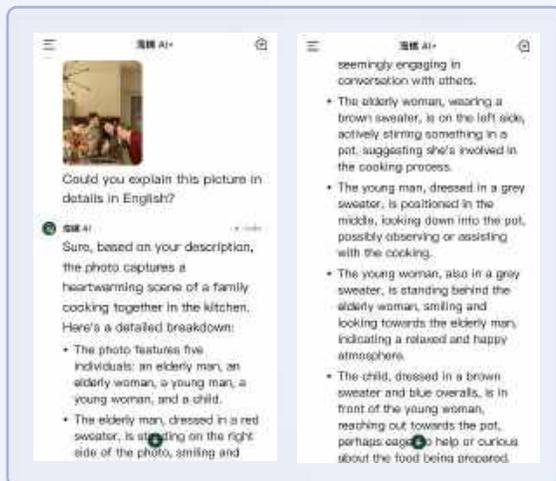


- The only Asian AGI company with a **global toC and toB footprint + revenue**
- ToC App: Multiple super apps with an **average daily user time spent of 100 minutes+**, and the **daily total time spent globally is second only to that of ChatGPT**
- ToB Open Platform: **One of the largest in APAC in terms of external query volume with 2,000+ paying customers**, and has started expanding into Southeast Asia and the Middle East

Only LLM Company in APAC with Full Multi-modality Models



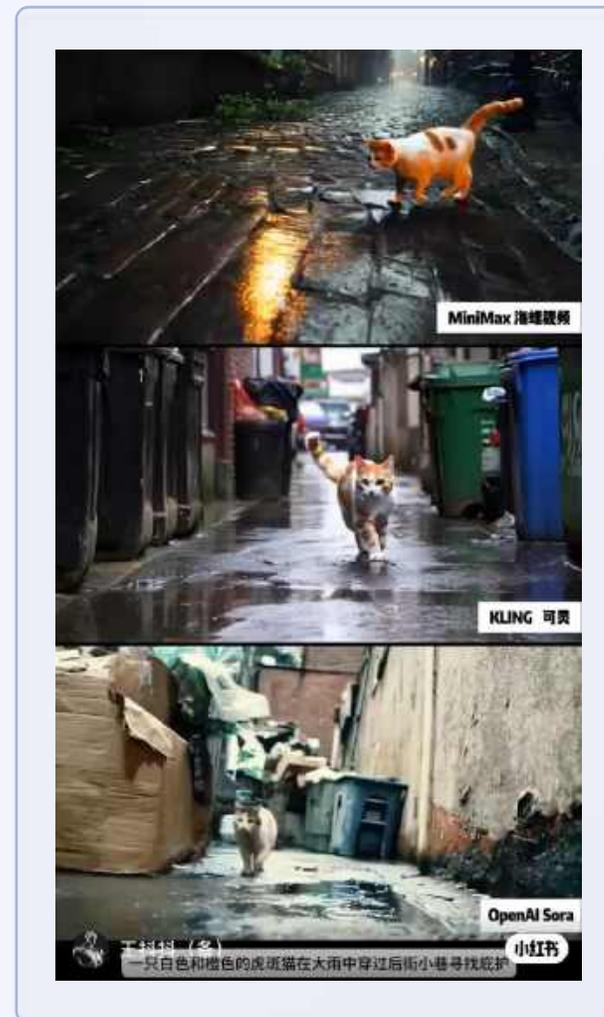
Aa Text



Audio/Music



Image/Video



MiniMax's Global AI Leadership & Publicity



Global Media

TECHNOLOGY | ARTIFICIAL INTELLIGENCE | Follow

One of America's Hottest Entertainment Apps Is Chinese-Owned

Chatbot in Talkie uses AI to allow users to converse with famous people or a virtual romantic partner

Through June, Talkie ranks No. 5 among the most-downloaded free entertainment apps in the U.S., according to Sensor Tower, a market researcher. That ranking puts it behind the likes of Warner Bros. Discovery's "Max," Netflix and Tubi.

More than half of Talkie's 11 million monthly active users are in the U.S., with popularity also strong in the Philippines, the U.K. and Canada. That puts Talkie within striking distance of the leader in the AI chatbot-companion category, Character.AI, run by an Andreessen Horowitz-backed firm in Silicon Valley, which has roughly 17 million monthly users, according to Sensor Tower.

According to The Wall Street Journal, **Talkie**, by MiniMax, has become one of the most popular entertainment apps in the United States. It ranks **5th in the list of top free entertainment apps in the U.S.** In addition to the U.S. market, it is also popular among users in **the Philippines, the United Kingdom, and Canada.**

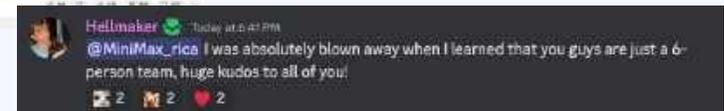


Hailuo AI, MiniMax's text-to-video product, has garnered significant attention and coverage from global media.

Sources: [SCMP](#), [VentureBeat](#), [Yourstory](#) and social media.



Social Media



MiniMax's text-to-video product, Hailuo AI, has received overwhelming positive feedback from users across global social media platforms, including X, Reddit, Discord, and YouTube.



Technological Foresight

- The most efficient technological route and innovation: front-runner **(No.1 in APAC, No.2 globally)** to launch the **Commercial MoE large-scale model**
- **Original Linear Attention algorithm reaches the international top tier**, solving the problem of computational complexity increasing with the square of data length, rapidly improving training and inference efficiency, **benchmarked to GPT-4o**



Full-stack Multimodality

- **The only startup in APAC** that possesses **text models, voice models, and visual generation and understanding models**
- Launching a multimodal model **similar to GPT-4o soon, incorporating the best text to video models(self-developed) currently on the market**, leveraging three rich modalities to create an immersive product experience



Low Inference Cost

- The inference cost of MoE **is 70% lower** than dense
- Strong cluster scheduling capabilities, with continuous optimizations such as **quantization techniques and storage caching**
- **Unified training & inference, real-time scheduling system** to balance the computational demand in peak and off-peak times



Efficient Large-scale Cluster Training Framework

- Customized large-scale computing cluster, tailored for efficient training of **trillion-parameter MoE model**
- **Self-developed training framework** with efficient experimental efficiency and parallel strategies, supporting dynamic cluster topology optimization, and deep **optimization for FP8 training**

Technological Foresight: Mixture of Experts (MoE) + Linear Attention



In the previous model Abab-6.5s, MoE is 3-5 times faster than Dense models



The first company in APAC to develop the next generation MoE+Linear Attention, efficiently achieving infinite length input and output, significantly improving training and inference efficiency



MiniMax-abab 6.5: Trillion-Parameter MoE Architecture LLM --> Minimax Text-01



Stronger performance (equivalent to GPT-4 in comprehensive Chinese capabilities)

Model	中文综合 AlignBench	英文综合 MT-Bench	知识 MMLU	基础算数 GSM8K	数学解题 MATH	逻辑推理 BBH	编程 HumanEval	指令遵从 IFEval
abab-6.5	7.97	8.82	79.5	91.7	51.4	82.0	78.0	79.6
abab-6.5s	7.34	8.69	74.6	87.3	42.0	76.8	68.3	74.6
GPT-4	7.53	8.96	84.6	92.0	52.9	83.1	67.0	85.6
GPT-3.5	6.08	8.21	70.0	57.1	34.1	66.6	48.1	70.5
Gemini 1.5 Pro	7.33	8.93	81.9	91.7	58.5	84.0	71.9	82.3
Claude 3 Opus	7.62	9.00	86.8	95.0	61.0	86.8	84.9	91.9
Claude 3 Sonnet	6.70	8.47	79.0	92.3	40.5	82.9	73.0	87.8
Claude 3 Haiku	6.42	8.39	75.2	88.9	40.9	73.7	75.9	85.4

Content Understanding

Quality inspection for call, information understanding & recommendation, intent recognition, etc.

Long Text Processing

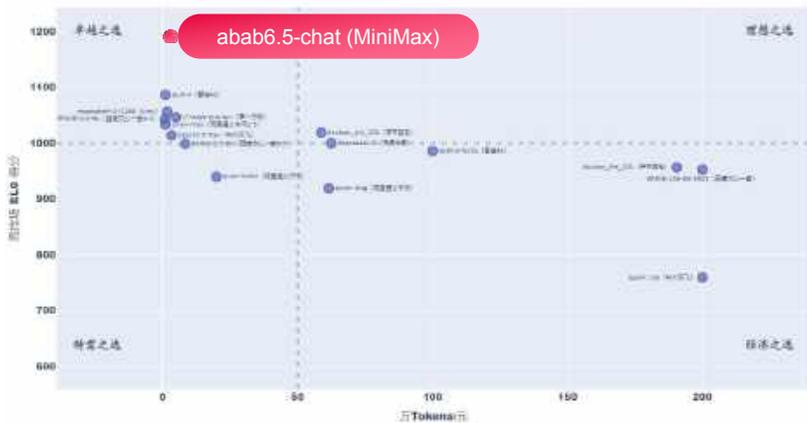
A context length of 245K, approximately 400,000 characters, can organize over 10 hours of meeting minutes and phone recordings in one go

Text Generation

Argumentative essays, academic papers, research reports, storytelling, etc.

Information Extraction

100% retrieval within 200K characters



Model	API Context Length (MAX)	Output Speed (tokens/s)	Input Speed (tokens/s)	TPM Limit
abab6.5s	245K	50~70	4000~5000	Unrestricted
Moonshot	128K	20	2000	Adjustable Based on Payment
Erniebot-4.0	8K	10~20	--	Separate Charge
Erniebot--3.5-128K	128K	35~50	--	Separate Charge
Qwen-Max	32K	15	2500	Application required
GLM-turbo	128K	40~40	3000	Adjustable Based on Payment
GLM-4	128K	25~35	1000	Adjustable Based on Payment
Baichuan4	32K	20	1200~1300	Not Allowed for Adjustment
Deepseek V2	32K	10	1000	Not Allowed for Adjustment

Music and Voice Model Examples

Music model: abab-music-1



Electronic



Rock and roll



Folk



Country Music



Pop Music



Realistic

Multi-genre

Artistic

Voice model: abab-speech-1



Cantonese



English



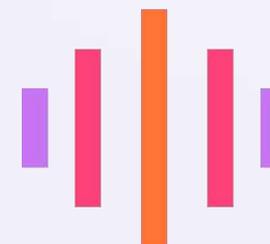
Spanish



Japanese



Korean



Emotional

Human-like

Multilingual

Hailuo AI: Key Technical Advantages



Text-to-video:
excellent instruction
following

Powerful
processing (camera
motions, complex
scenes, nuanced
expressions with
emotions)

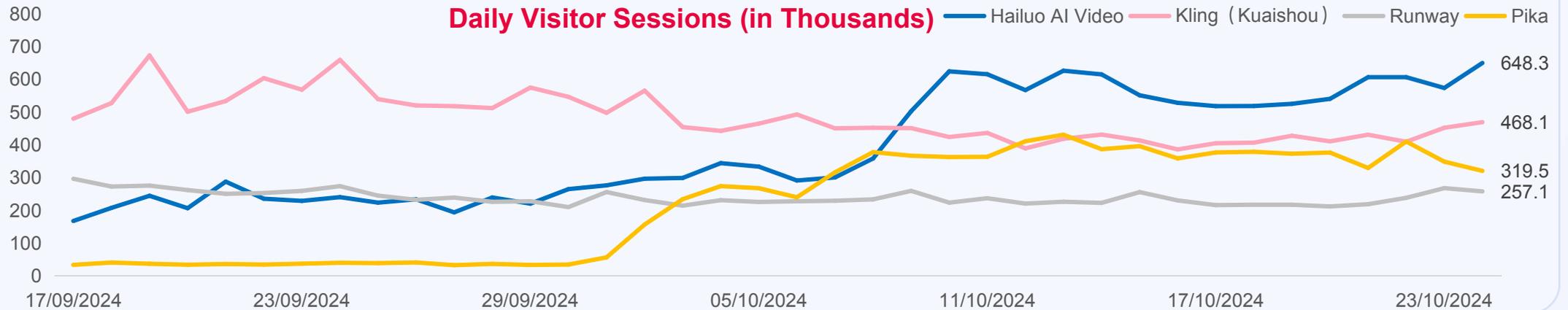


Diverse styles
(super realism;
Eastern; ACG; etc.)

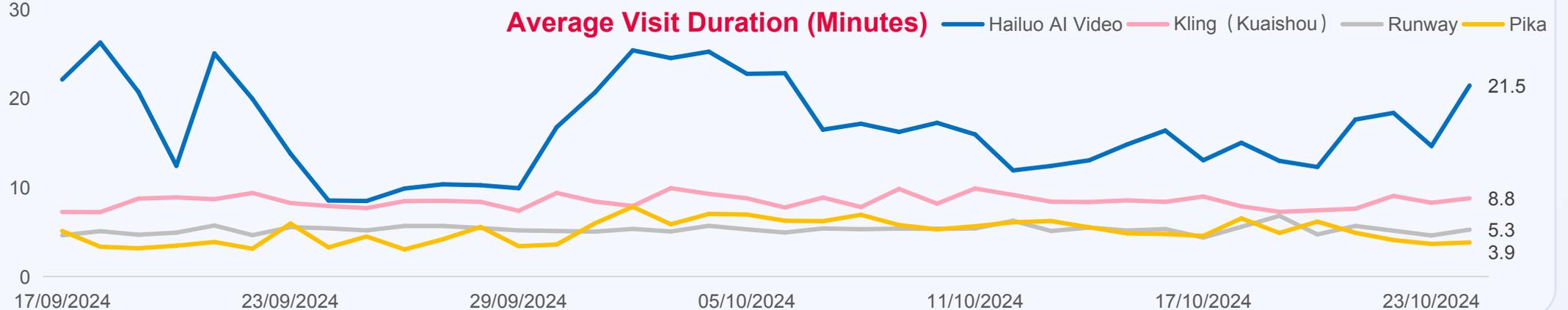
Both the Visitor Sessions and Average Visit Duration of Hailuo AI Video are Leading Among Its Peers



Without any promotion, the # of visitor sessions has rapidly surpassed that of Kling, Runway and Pika



The average visit duration is twice that of Kling and three times that of Runway and Pika



Source : SimilarWeb as of Oct.24, 2024; The Hailuo AI PC domain names include hailuoai.video and hailuoai.com.

The Real Answer: Technological Superiority Unleashes Creativity



Model Name	Source	Total Score	Quality Score	Semantic Score
MiniMax (abab-video-1)	VBench-Long Team	83.41%	84.85%	77.65%
Runway Gen-3 (2024-07)	VBench-Long Team	82.32%	84.11%	75.17%
Shanghai AI Lab Vchitect-2.0 (VEnhancer)	VBench-Long Team	82.24%	83.54%	77.06%
Kuaishou Kling (2024-07 high-performance)	VBench-Long Team	81.85%	83.39%	75.68%
Zhipu CogVideoX-5B	VBench-Long Team	81.61%	82.75%	77.04%
Shanghai AI Lab Vchitect-2.0-2B	VBench-Long Team	81.57%	82.51%	77.79%
Zhipu CogVideoX-2B	VBench-Long Team	80.91%	82.18%	75.83%
HPC-AI tech OpenSora_V1.2	VBench-Long Team	79.76%	81.35%	73.39%
PKU-YuanGroup OpenSoraPlan_V1.1	VBench-Long Team	78.00%	80.91%	66.38%
HPC-AI tech OpenSora_V1.1	VBench-Long Team	75.66%	77.74%	67.36%
Tencent Mira	VBench-Long Team	71.87%	78.78%	44.21%

Note: VBench-Long, a comprehensive benchmark suite for long video (5 seconds and above) generative models. VBench-Long carefully decomposes video generation quality into 16 comprehensive dimensions to reveal individual model's strengths and weaknesses.

Comparison – Complex Instructions

MiniMax - Hailuo



- Hailuo's **Split Screen** allows for more artistic expression
- **Complex and long instructions** - Hailuo accommodates for up to 2,000 words, while peers only do 500-1,000 words.
- **Superior instruction following** – Hailuo can cover 80%+ of complex instructions with great logic.
- Hailuo can achieve **multiple styles changes** within one clip
- Hailuo has good **physical rules** and can achieve complex actions

Runway Gen-3



***Prompt:** The scene is divided into two parts, left and right, separated by a boundary. On the left is a quiet living room with a soft sofa and a lazy puppy dozing off. On the right is a stormy sea with a small boat bobbing in the rough waves. The styles on both sides are very different. Suddenly, the waves rolled in and the sea crossed the dividing line, splashing down on the living room floor and waking the puppy. The puppy froze for a moment, then looked curiously at the water stretching from the right.*

Showcases from Users – Revolutionising the Way of Video Creation



MV - I Feel Your Pain



- Hailuo helps artists in visualizing their music through **intricate scenes, consistent characters, and nuanced expressions, all brought to life instantly**

Short Video - Avalanche



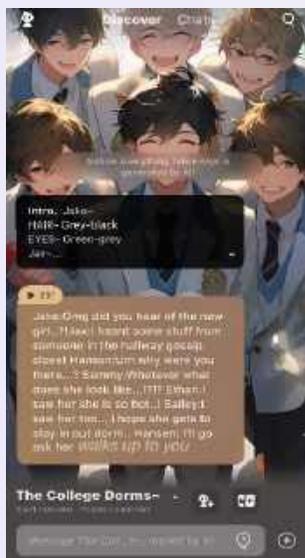
- Hailuo empowers every user to unleash their creativity, **enabling everyone to become an artist without incurring substantial costs**

A Series of Super Apps + Internationally Recognized ToB Open Platform



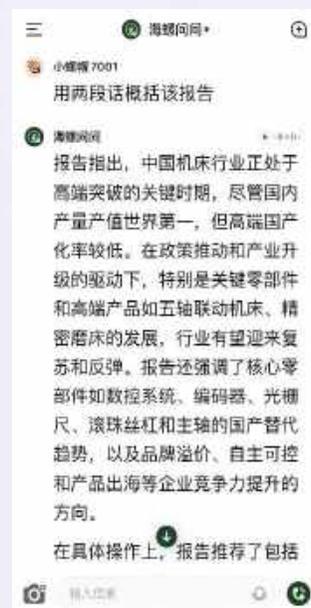
- **ToC: Talkie/Xingye is the No.2 AGI entertainment app globally with over 60M registered users**
- **ToB: one of the largest AGI open platform in APAC**
- **Processing over 5 trillion of text tokens daily, reaching 50% of OpenAI's, besides with 25M+ images and 600K+ hours of audio generated daily**

Talkie and Xingye



Average daily time spent > 100 minutes;
Ranked top 10 in the entertainment category in 10+ countries' app stores

Hailuo AI



Leading in video/music generation, voice call function, and real-time searching

Video + Music



Global best-performing text-to-video and text-to-music platforms

Open Platform Selected Customers



Xiaohongshu

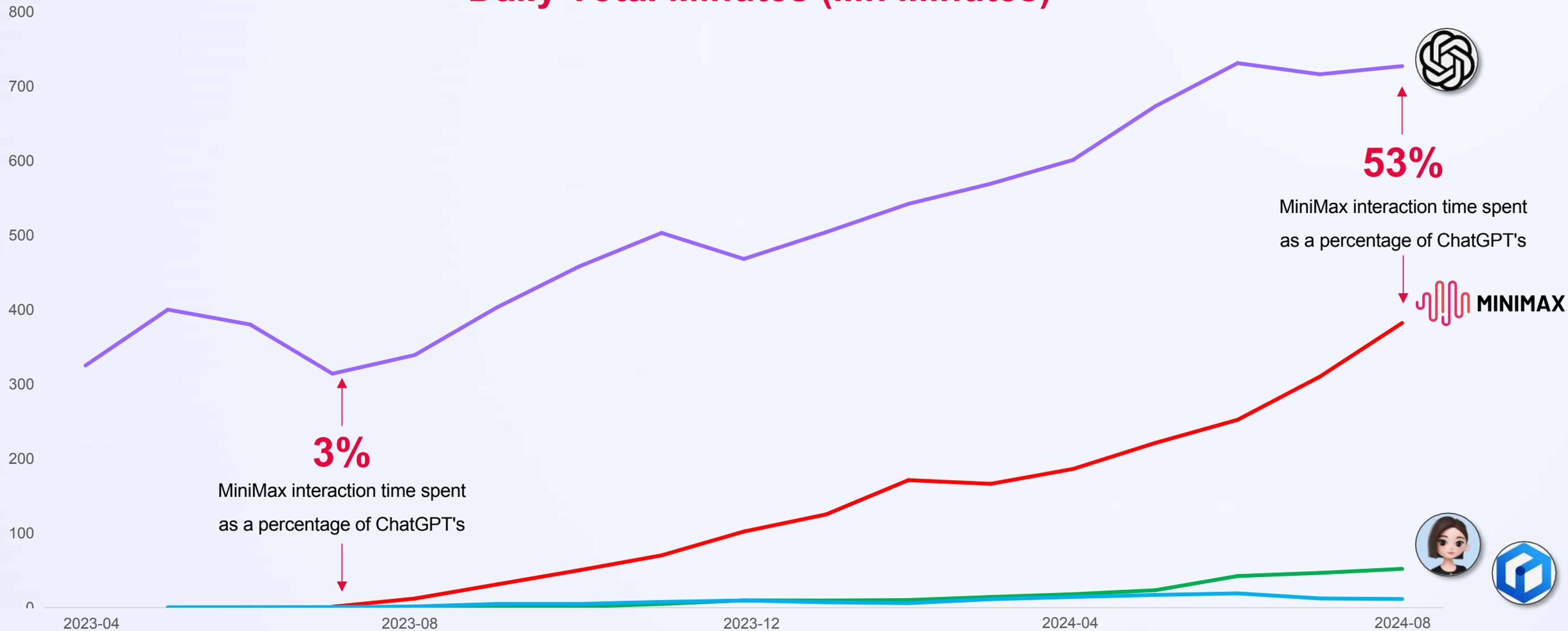


Have expanded to international KA customers

Interaction Time Spent Grew from 3% of ChatGPT's to 53% in Only One Year



Daily Total Minutes (Mn Minutes)



Sources: Quest Mobile, Mobile data from Sensor Tower, Web data from Similar Web & YipitData

AGI Interaction League Table



Daily Total Minutes (Mn Minutes)



Sources: Quest Mobile, Sensor Tower, Similar Web as of 2024-07-22 ~ 2024-07-28

MiniMax has 60M+ Users Globally: Our Models are Widely Used by Users and Enterprises Around the World



Open Platform: Empower Enterprises and Developers with Large Model Capabilities



Tourism

- Itinerary Planning
- Tour Assistance
- Travelogue Writing
- Customer Services Handling for Ride-hailing



Intelligent Hardware

- Intelligent Hardware
- Smart Home
- Online Shopping Assistant
- Voice Assistant for Mobile Phone



Retail & E-commerce

- Product Planning
- Live Broadcast Planning & Assistant
- Person-to-goods Matching



Office Efficiency

- Excel Processing
- Text and PPT Generation
- Enterprise Q&A
- Meeting Minutes & Task Assignment



Education

- Master Replication
- Oral Practice Companion
- AI Assistant
- Course Production



Internet Related

- Entertainment & Social Networking
- Search & Recommendation
- CV Matching
- CV & JD Optimization



Game

- Intelligent NPC
- Script-based Role Play Game
- Text Adventure
- Game Assistant



Finance & Insurance

- Investment Consultant
- Confidant
- Reliable Sales Assistant
- Star Product Sales



Healthcare

- Consultation Assistant
- Medication Recommendation
- Follow-up Conversation
- Medical Knowledge Base Search



Intelligent Automobile

- Casual Conversation Companionship
- Car Control
- Intelligent Manual
- Light Game Entertainment



WPS & MiniMax | Improve Work Efficiency by Generating and Processing Ultra-long Texts

PDF 文档 
可快速提炼重点，
依据内容，回答问题。

 WPS AI
上新了。



PDF: content extraction, summarization and Q&A

PPT 演示 
也能自动生成
你想要的内容。

 WPS AI
上新了。



PPT: full deck generation on a given topic

- 100+ WPS APIs are supported by MINIMAX
- Generate abundant and compliant contents

WPS & MiniMax | Automatic Excel Processing and Analysis with Complex Formulas Calls

Excel: formula generation, data analysis and chart generation

The screenshot shows the WPS AI interface with a data table and an AI chat window. The table contains columns for product categories, prices, costs, and sales. The AI chat window is open, displaying a welcome message and options for document processing, analysis, and chart generation.

商品大类	商品小类	建议售价	单位利润	采购总成本	平均成本	单位利润率	商品采购入库数	总销售量	剩余库存数	缺货报警	总销售成本	总销售额	总利润
家用电器	洗衣机	¥950.00	¥200.00	¥120,000.00	¥750.00	21.05%	160	152	8	库存数不超	¥114,000.00	¥144,400.00	¥30,400.00
家用					¥1,990.00	23.46%	230	51	179		¥101,490.00	¥132,600.00	¥31,110.00
家用					¥1,700.00	14.14%	50	44	6	库存数不超	¥74,800.00	¥87,120.00	¥12,320.00
家用					¥2,000.00	42.86%	160	46	114		¥92,000.00	¥161,000.00	¥69,000.00
电脑					¥900.00	18.18%	300	25	275		¥22,500.00	¥27,500.00	¥5,000.00
家用					¥2,000.00	49.99%	200	21	179		¥42,000.00	¥83,079.00	¥41,079.00
家用					¥1,900.00	51.27%	60	35	25		¥66,500.00	¥136,465.00	¥69,965.00
手机					¥100.00	64.16%	400	230	170		¥23,000.00	¥64,170.00	¥41,170.00
电脑					¥584.00	12.31%	500	25	475		¥14,600.00	¥16,650.00	¥2,050.00
手机					¥3,000.00	38.78%	300	53	247		¥159,000.00	¥259,700.00	¥100,700.00
电脑					¥2,000.00	33.31%	400	83	317		¥166,000.00	¥248,917.00	¥82,917.00
手机					¥50.00	36.71%	500	119	381		¥5,950.00	¥9,401.00	¥3,451.00
电脑					¥1,800.00	22.35%	300	48	252		¥86,400.00	¥111,264.00	¥24,864.00
手机					¥1,760.00	41.31%	400	57	343		¥100,320.00	¥170,943.00	¥70,623.00
手机					¥6,000.00	22.07%	80	4	76		¥24,000.00	¥30,796.00	¥6,796.00

标记重点

- 4,000+ formulas are automatically called
- Up to 80 layers of nested formula calls
- Formula generation reinforcement learning environment

通过聊天

表格海外版

Tencent Meeting & MiniMax | Complete Complex Tasks such as Information Extraction, Content Analysis, and Intelligent Reminders



- Automatically schedule meetings
- Summarize meeting contents
- Summarize speaking viewpoints
- Organize meeting action items
- Search for content specified by members
- Automatically assign tasks

Xiaohongshu/Red & MiniMax | Combine Searching for High-quality Human-generated Content to Address Hallucination in Large Models

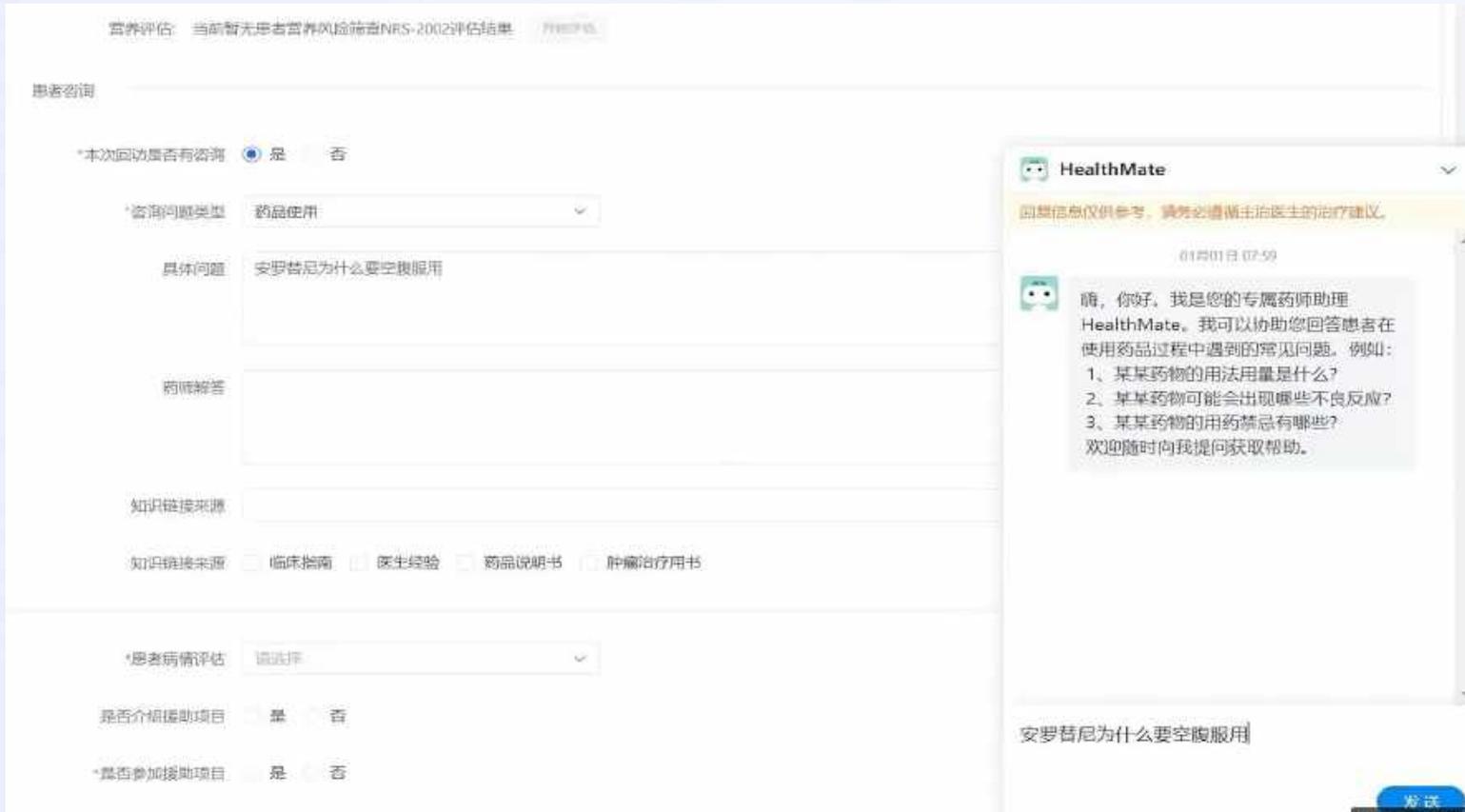


- Query term auto-expansion + search result summarization + generation based on both query and search content
- The first to launch large models + search in China
- 95% accuracy in model responses, surpassing Microsoft Bing-chat

Cowell Health & MiniMax | LLM + Vector Database for AI Pharmacist Assistant



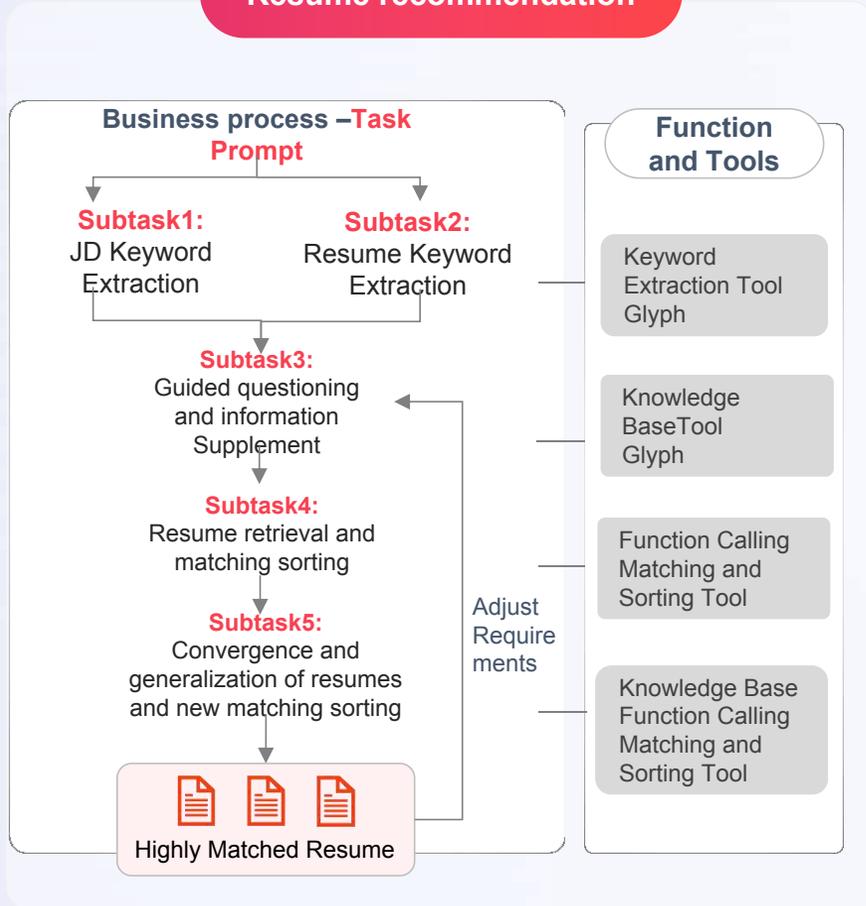
Cowell Health is a leading drugstore chain with 10,000+ pharmacies in 20+ provinces in China



- Create vector embeddings with historical information automatically
- Generate professional responses with LLM + vector search
- Outperform average pharmacist

Zhaopin & MiniMax | Multitasking Agent Boosts Resume Matching on Recruiting Platform

Resume recommendation



AI recruitment solutions for all scenarios

General scenarios

- Resume Diagnostics** - Optimize resume structure and content, highlight user's professional skills and achievements.
- Resume Writing** - Write a professional resume to help you stand out in the job market
- Career Planning** - Develop a strategic plan for the user's career development
- Interview Guide** - Provide interview techniques and strategies to enhance the success rate in interviews.

Specific Scenarios

- Salary Inquiry** - Based on industry data, help users to negotiate a better salary
- Career Assessment** - Through professional assessments, providing a scientific basis for career choice.
- Job Recommendations** - Recommend high-match jobs based on users' skills and help users find ideal job faster.

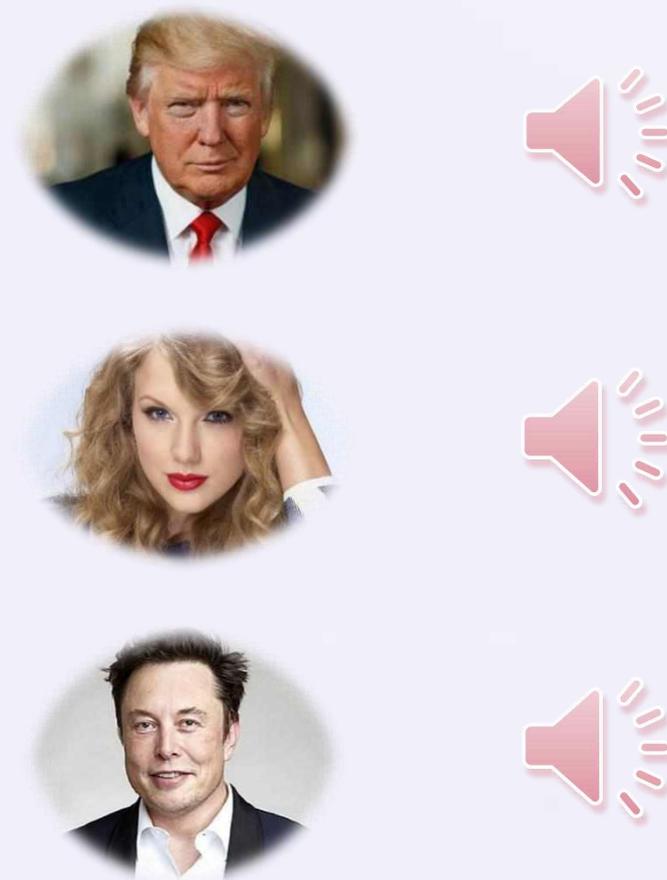


- Recommend the most suitable resume for employers
- Provide resume diagnosis, resume writing, career planning, interview guide, salary inquiry, career assessment, and job recommendation services for job seekers

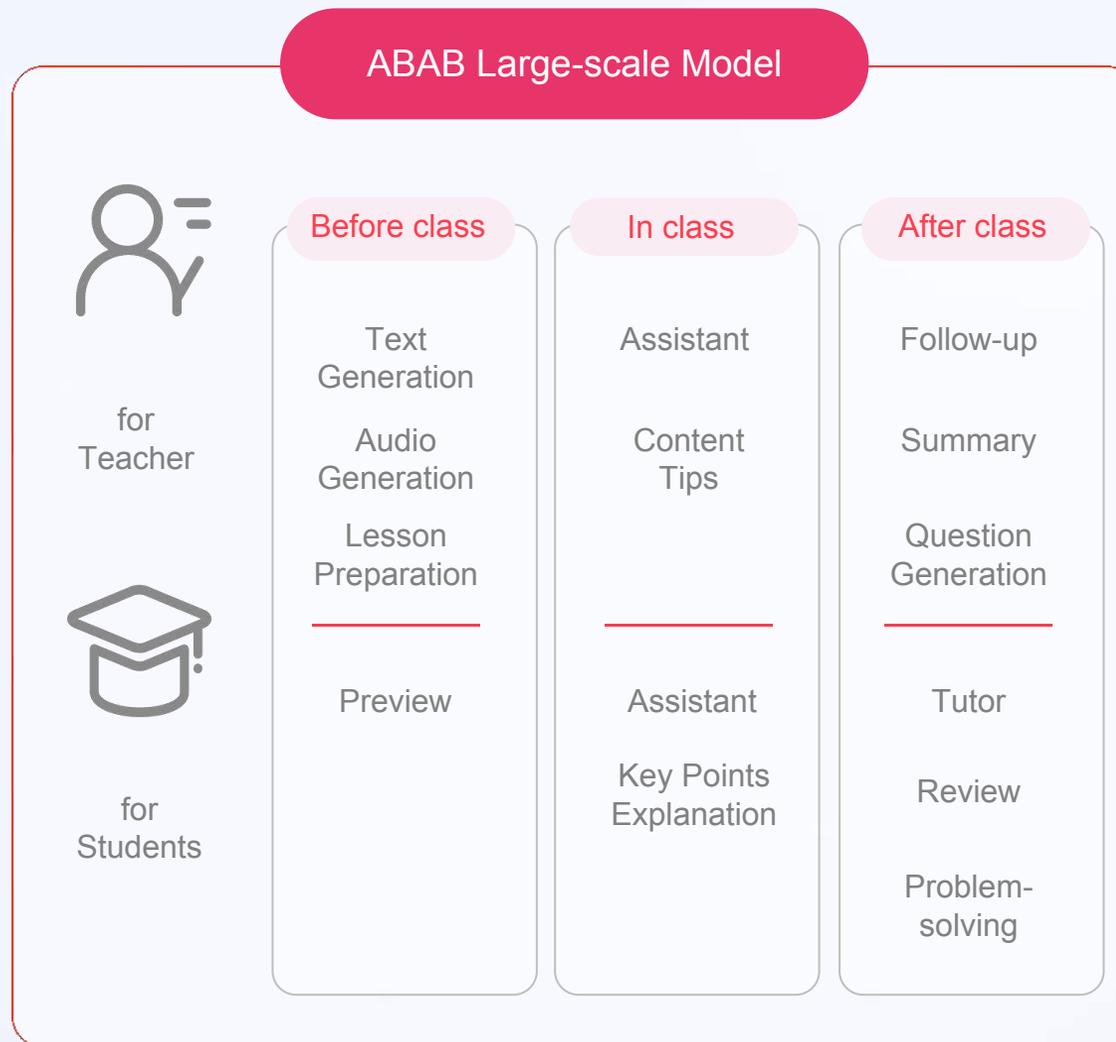
Qidian & MiniMax | Audiobooks

7,000+ complete audiobooks, with all top serial novels audio-enabled

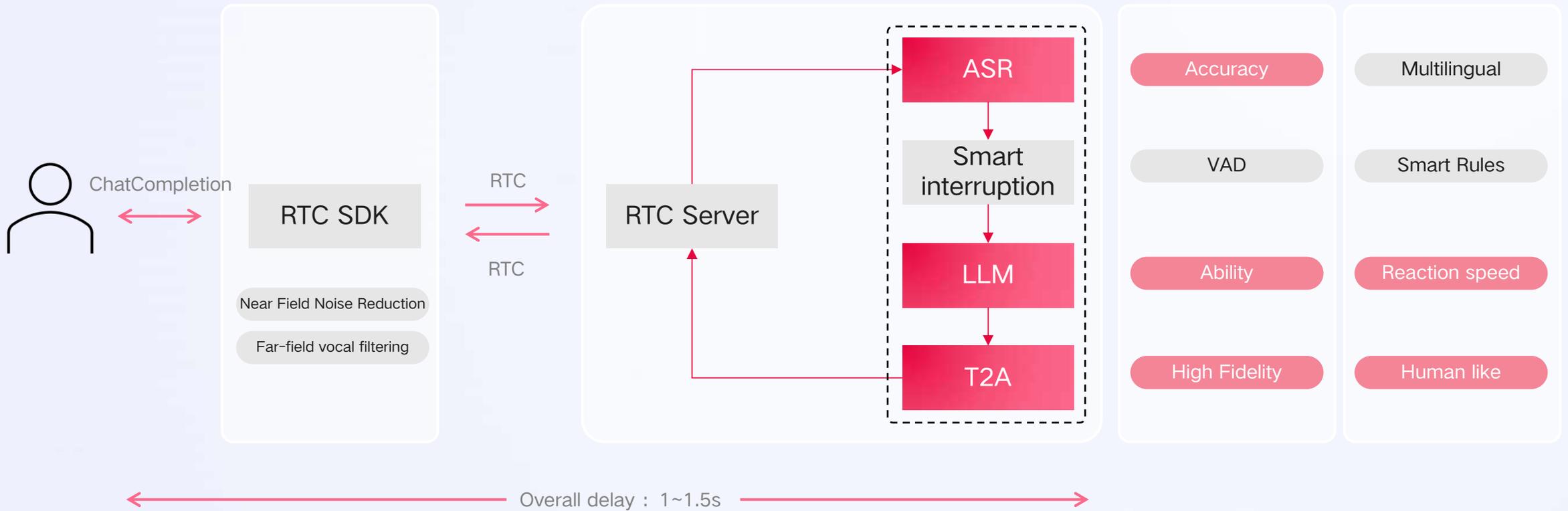
High fidelity, hyper-realistic, and highly scalable



Gaotu & MiniMax | Replicate the Voice Timbre of Outstanding Teachers and Create Online Courses



Intelligent Conversation – Realtime API



- * The ability to provide VAD in RTC mode
- * MM-ARS currently only supports Chinese and English.
- * This API provides two interface methods: WebSocket & HTTP

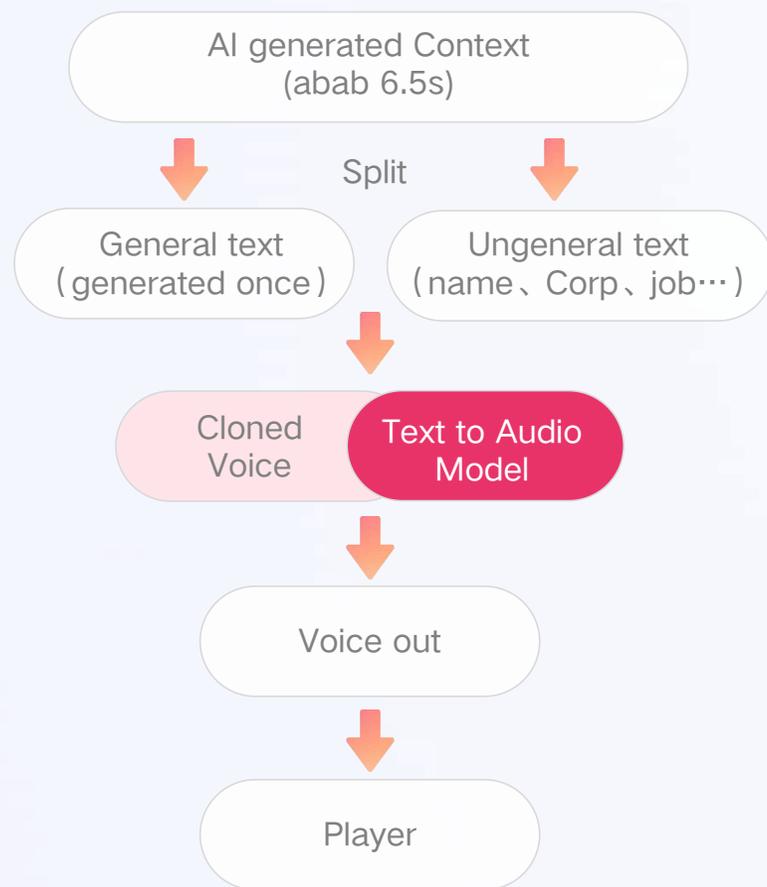
MM hold

RTC or User hold

Liepin & MiniMax | Intelligent outbound calling to achieve business reach



WorkFlow



Example



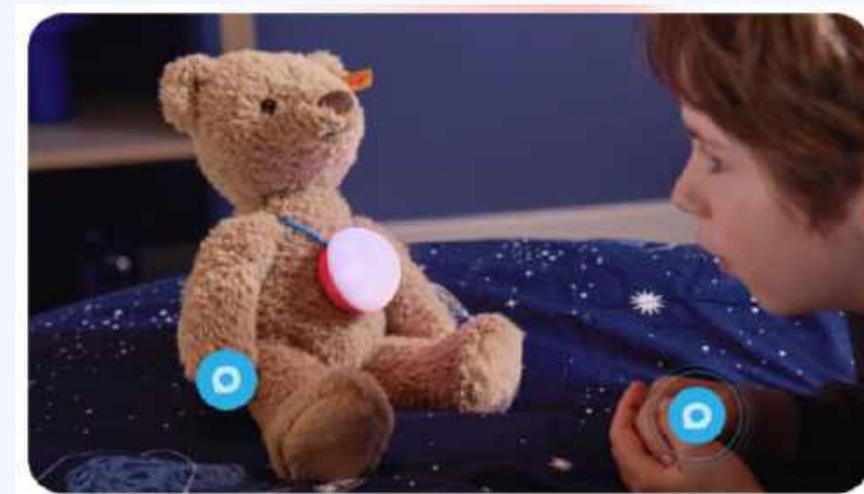
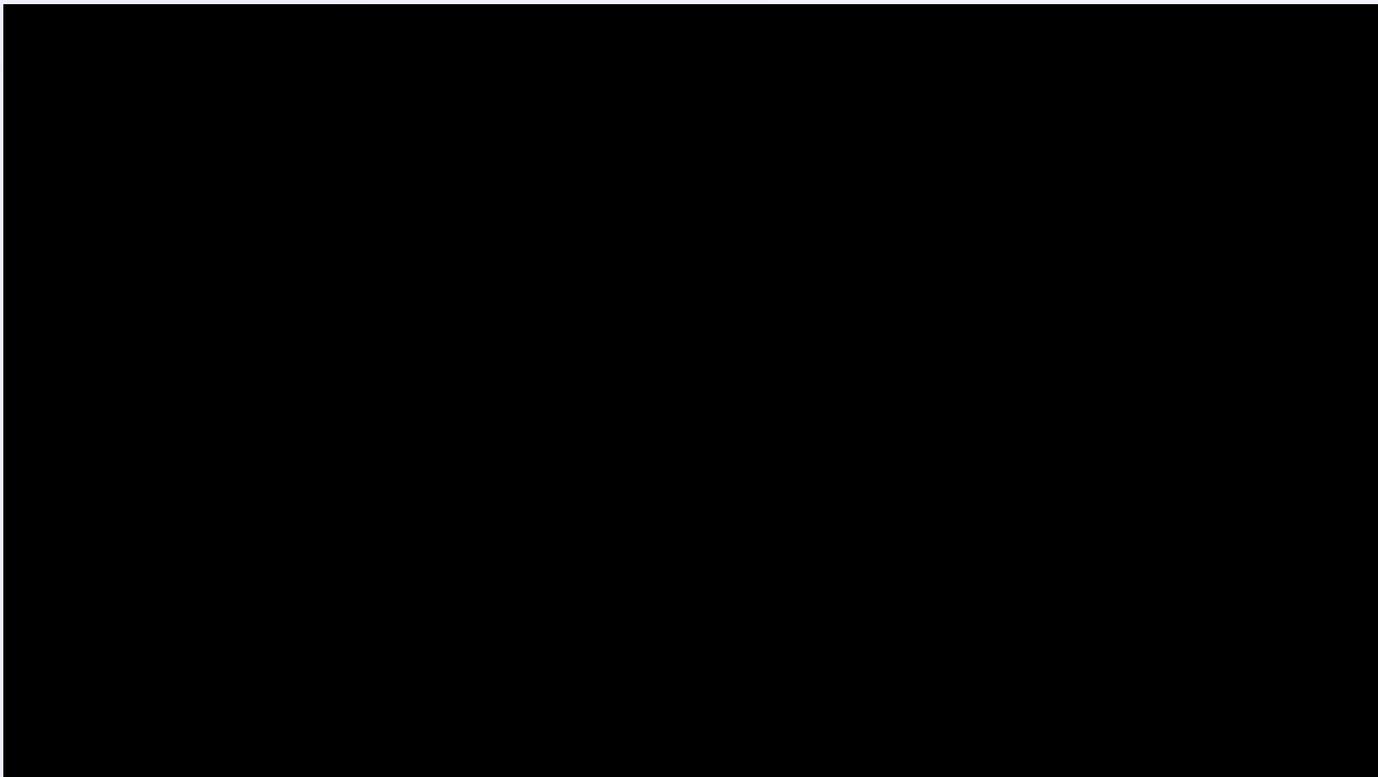
Core Case

Simulating real human communication, natural voice significantly improves customer call acceptance rates and satisfaction, promoting more positive customer interactions.

The human natural voice model can synthesize a variety of different tones, speaking speeds and emotions according to the conversation context, better adapting to the communication styles of different customers.

The text-based model can understand the user's voice content and better answer user questions

HAIVIVI & MiniMax | Creating a Child Growth LPA



- Immersive role-playing
- Hyper-realistic voice synthesis T2A
- Long-term memory recording of children's growth process
- Rich content including stories, oral dialogues, encyclopedic knowledge, etc.
- Parents can customize roles, input expected dialogues, set restrictions
- View growth records, understand children's emotional development, and timely detect their psychological development status



Overview

Customer
Pain Points

High customer service costs and poor user experience

Our
Solution

Automatically categorize and label user complaints, determine if manual intervention is required or not, reduce labor costs, and improve customer service efficiency

Labels

Opinion Type

Complaint, roast, consultation, etc.

Business Line

Express, Premier, Luxe, Taxi Hailing, Hitch, third-party services, etc.

User identity

Drivers, passengers, third-party, journalists, etc.

Emotions

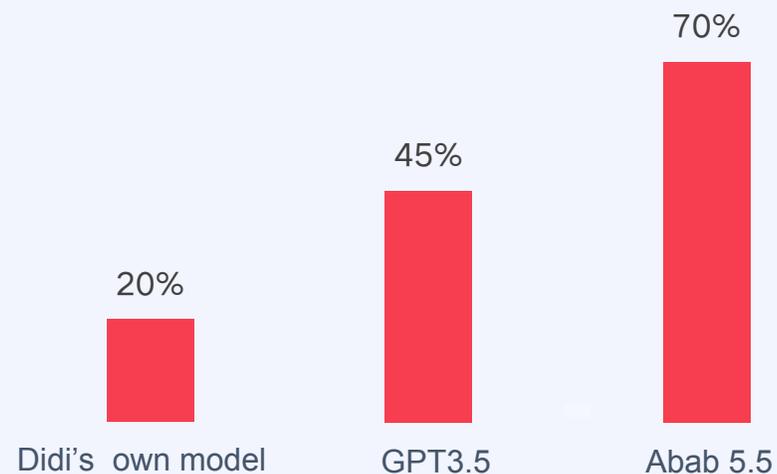
Emotional, angry, calm, etc.

Processing
method

No action, coupon, operational activity, manual intervention, etc.



The accuracy is significantly improved



Kuaishou E-commerce & MiniMax | Empower Pre-broadcast Planning, Live Broadcast Recommendations, Customer Services, and Reviews



Sentiment Analysis

User comments and user danmaku content are labeled and categorized.

Customer Service

Audio large models replace staff for after-sales information reminders, event reminders, and logistics information reminders.

Store Management

Quickly generate store inspection reports and intelligently analyze and mine various data of the store.

Live data analysis

With the ability of function calls, complete the review and data organization of the entire live streaming process.



- AI assists in part of human tasks to achieve cost reduction and efficiency improvement.
- Mid-term goal: Achieve coverage of more than 50% of routine and repetitive tasks in business processes.
- Final goal: Complete the handover and monitoring of the entire AI business process.

FWD Group & MiniMax | Insurance Agent to Help Brokers Improve Efficiency and Customer Satisfaction



Create an insurance avatar for the insurance industry, providing professional advice and customized product recommendations.

Scenario Requirements

Selection Process

Customers often find it difficult to understand the specific differences and coverage of each product, making it impossible to make the most suitable choice for their needs.

Knowledge Popularization

Ordinary customers often lack sufficient insurance knowledge and may not grasp the significance of insurance or how to select the appropriate insurance product.

Claims Process

When an insurance incident occurs, customers often feel confused and frustrated because they are not familiar with the claims process and the required materials.

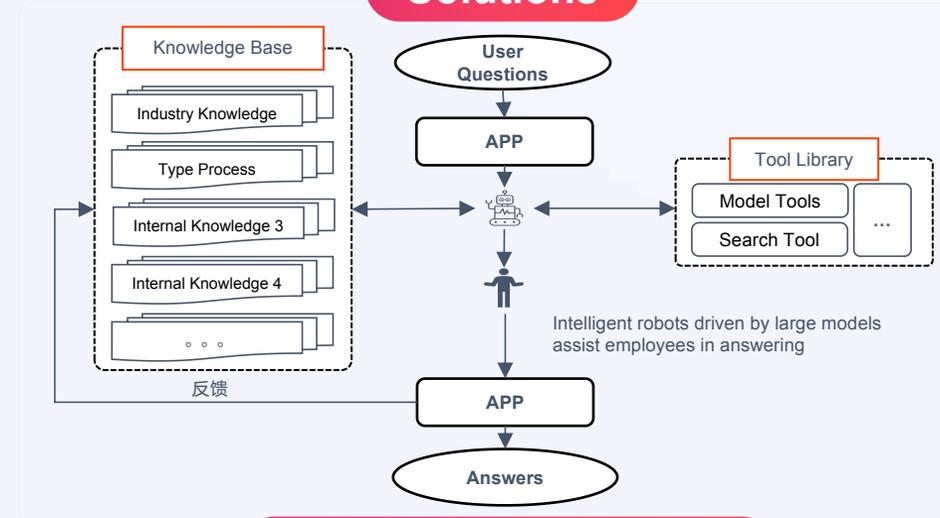
Personalized Service

Traditional insurance services often fail to provide personalized advice tailored to the unique needs of each customer, leading to a subpar experience.

Service Timeliness

When customers need help, traditional service channels (such as telephone, email) may not respond in time, affecting satisfaction.

Solutions



Business Objectives

Professional Consultant

Provide comprehensive and optimal professional advice on investment, tax plans, education and medical care, retirement and life planning.

Star Salesperson

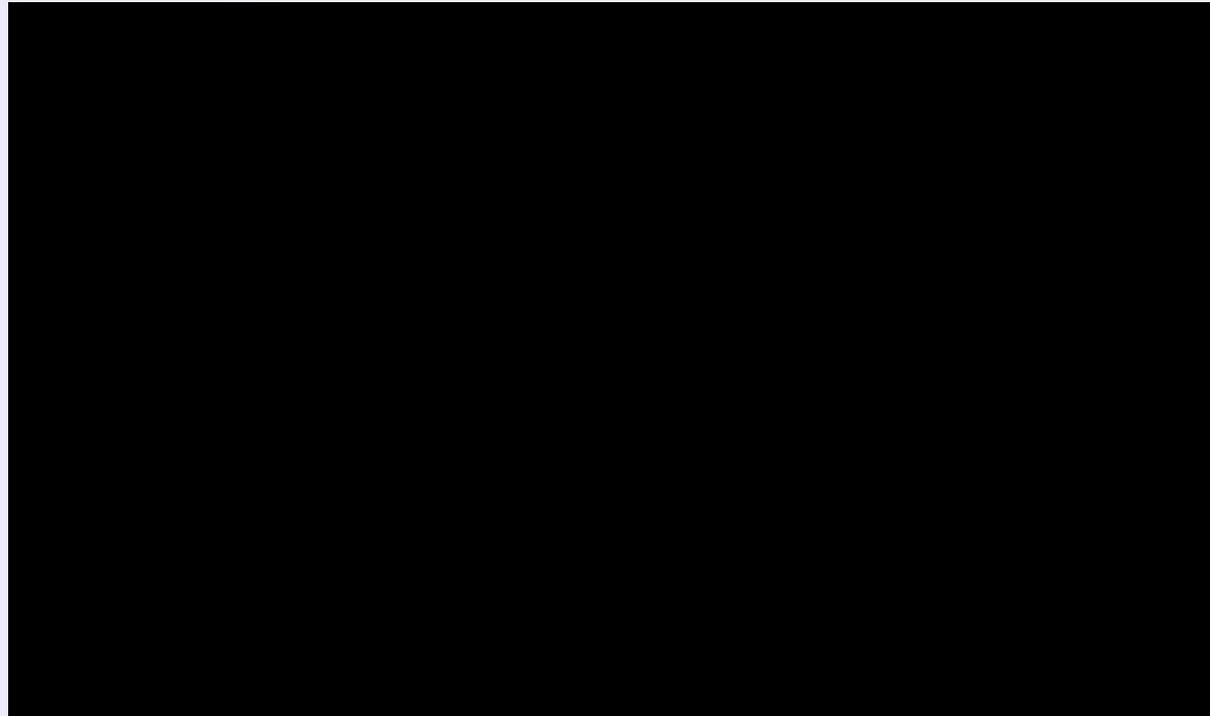
Recommend products tailored to the customer in the most personalized way to improve the transaction rate.

Reliable Salesperson

With profound knowledge of product details, accurately answer customers' questions about various stages of products and company details.

Best Confidant

Online 24/7, provide customers with companionship and emotional value, deepening customer trust.



- Billions of High-Quality Dialogue Data
- Efficient Intelligent Retrieval
- Use of Large Model to Synthesize Voice, Full of Emotion
- 30s High-Fidelity Voice Cloning

Companies That Can Realize a Close Loop of Both Technology and Applications will Gain Maximum Commercial Value in the AGI Era

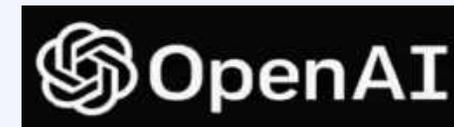


AI Technology
Evolution

Search

Recommendation

AGI



Thanks!