

Tencent 腾讯 | CSIG
云与智慧产业事业群

大模型下 腾讯云AI布局与能力实践

2026年3月 | 成都



腾讯近况：持续加大AI布局，25年AI大模型相关投入超1000亿

2025全年绩摘要

全年总收入

7518 亿元人民币

同比增长 14%

全年利润

2807 亿元人民币

同比增长：18%

增值服务业务

3693 亿元人民币

同比增长16%

营销服务业务

1450 亿元人民币

同比增长19%

金融科技及企业服务

2294 亿元人民币

同比增长8%

AI基础设施：百万节点，GPU投入数百亿

- 基础设施布局：腾讯全球服务器国内云厂商中最早过百万，在全球26个地区，建设超过70个可用区
- GPU采购：23年底公开数据，腾讯采购5万块H100 GPU，全球第三，国内第一。24年，25年腾讯在GPU采购也均排名国际国内前列。
- AI持续投入：2025年腾讯不断加大AI投入，25年AI相关投入超1000亿。

自研的AI：加码投入，筑造强引擎

混元大模型：万亿级参数规模，2025年发布发布了超过30个新模型，并开源多个模型

混元语言类模型
HY-2.0-Think

混元生图模型
HunyuanImage
3.0

混元生视频
HunyuanVideo
1.5

3D游戏创作
混元3D 2.1

增效的AI：助攻产业，高质提效能

腾讯广告

搜索广告
收入翻倍



视频号
广告收入同比增长60%

腾讯游戏
Tencent Games

游戏制作
大量采用AI技术

企业微信

深度应用AI技术
收入翻倍

腾讯会议
Tencent Meeting

腾讯云智能：离产业最近的AI

应用场景



产品全景图



腾讯云智能

8000+项

AI相关专利

1000+篇

论文入选全球顶会/期刊



IDC
亚太第一 中国第一
 IDC发布的《2024年AI公有云市场份额 [包括CV、对话式AI、自然语言处理、智能语音、机器学习]》，腾讯云位居第一梯队。
 入选IDC《MarketScape: 2022全球通用计算机视觉厂商评估》。中国唯一入围该报告的厂商，位于市场“主要玩家”位置。



Gartner
唯一连续5年入选
 入选Gartner魔力象限《2024 全球云 AI 开发者服务魔力象限报告》。唯一连续5年入围的中国云厂商；计算机视觉中国第二。



CAICT
 中国信息通信研究院
信通院产品能力最高等级认证
 腾讯云TI平台获得信通院人工智能开发平台功能评测的最高等级认证4级 [国内首批]，在全能力域的产品能力达到国内最高水平。

目录

1. AI基础设施
2. AI通用模型
3. AI大模型应用

01

AI基础设施:云与GPU算力

——技术先进与性价比首选

腾讯云，全球领先的云厂商

——遍布全球五大洲26个地区+70个可用区+丰富完善的云产品体系，服务于千行百业

基础服务：融合创新的算力底座

算力融合，要充分考虑智算与通算如何融合，服务于整体业务，让智算算力不是特立独行的存在，同时也需要考虑网络、存储等高性能基础设施。

算力选择，除了优先考虑国产化，从使用场景也要考虑其他算力融合，构建更适合于服务当下和未来的架构。

建设模式，可以考虑多种多样的模式，比如公有云API服务、公有云资源、分布式云等等，都有其优势，满足不同场景下的需求。



全网超 **100+w** 台服务器部署规模，算力调度总规模超过 **1.5 亿核**，并提供超 **16 EFLOPS*** 的智算算力

- 腾讯云灵活多样的云平台提供方式，可更好地协助政、企客户快速搭建\使用云资源。全球26个地区的资源分布、丰富的跨境业务服务经验，可助力区内企业扬帆出海。
- 面对需要数据不出域的客户，腾讯云可以提供分布式云和专有云模式，助力用户更灵活构建基础支撑服务。
- 一云多芯，支持CPU\GPU\NPU等多种资源集中管理，多卡混跑，**首家支持Intel、AMD、NVIDIA、Xilinx和Arm处理器的“多芯算力”提供商**

腾讯拥有全面领先的产品和技术

腾讯云计算

第一

AMD+100G网络, 行业性能

3000万+

峰值 QPS

腾讯云存储

第一

中国云存储解决方案市场
国内厂商排名

99.995%

服务可用性

100GB/s

带宽支持

腾讯云网络

国内首家

支持QUIC网络传输协议的
云计算服务商

99.99%

云联网可用性

腾讯云音视频

第一

连续6年视频云解决
方案市场排名

300ms

通话场景国际链路端
到端平均时延<300ms

50%+

相同视频主观质量下
节省50%+带宽及
存储成本

腾讯云数据库

世界第一

腾讯云数据库TDSQL登顶
TPC-C榜

8.14亿笔

交易/每分钟

70%

服务中国
Top 10银行的比例

腾讯云智能

第一

连续4年
“计算机视觉应用”
方面中国厂商排名

5000+项

AI相关专利

800+篇

论文入选
UCA/CVPR/ICASSP/ACL
等全球顶会/期刊

腾讯云原生

国内第二

入选Gartner魔力象限
报告挑战者象限
执行力维度排名

287%

客户3年投资
回报率

65%

极致资源利用率

腾讯企点

第一

中国CRM
市场份额

领导者类别

中国客户
数据平台 [CDP]

100万+

服务的
企业数量

腾讯会议

连续2年

入选Gartner《视频会议
解决方案市场指南》
代表厂商

4亿+

用户

220+

服务的国家
和地区

25亿次

会议协同

腾讯安全

大满贯

全球七大权威机构
测评

10亿级

为海量用户
提供安全防护

1500+

云安全技术
专利申请量
位列行业第一

1.5万亿次

年拦截攻击量

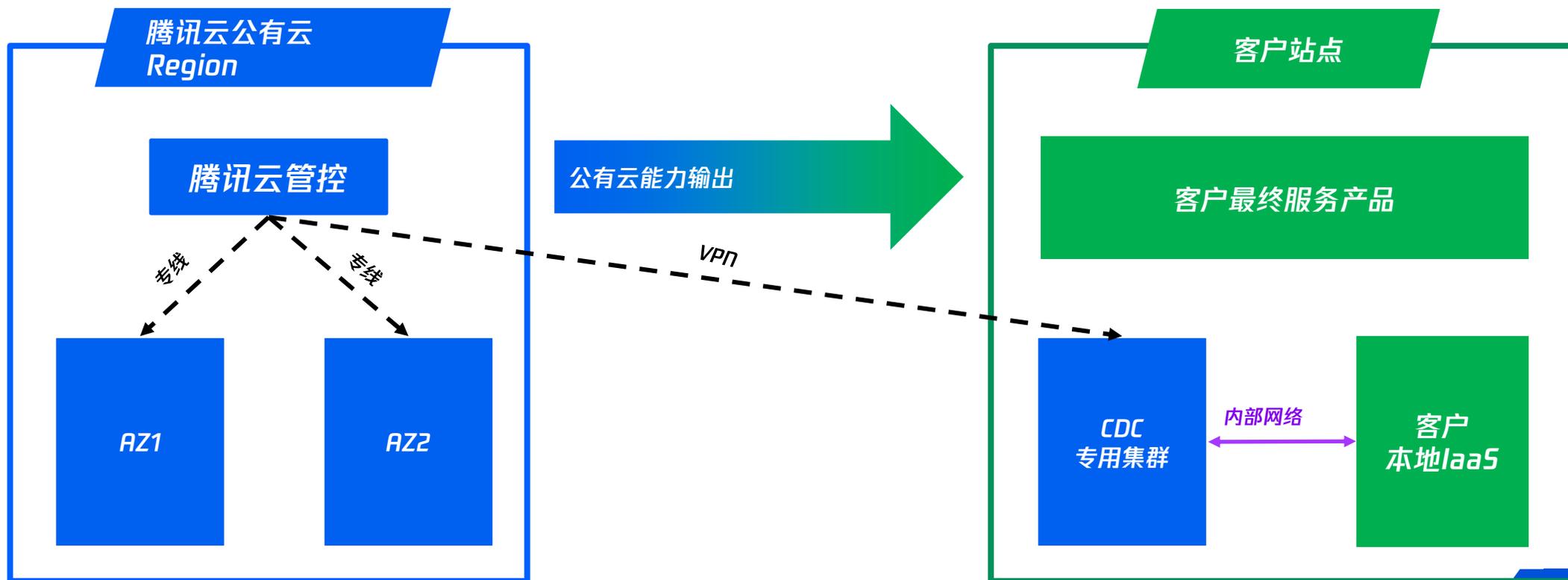
本地专有集群：建设在企业本地机房的专有云

将公有云延伸到客户机房

- 近场服务：用户选择部署位置
- 中心纳管：一致体验，混合同构管理
- 无需值守：软硬件由腾讯进行运维

云计算与本地IDC优势的极致融合

- 公有云的：敏捷弹性、丰富服务、安全可靠
 - 本地IDC的：超低时延、数据驻留
- 本地专用集群将两者优势合二为一



成为千行百业值得信赖的数字化助手



※ 更多产业解决方案及标杆案例, 详见《腾讯产业互联网》相关资料

02

AI通用模型：混元大模型

腾讯混元向智能的深度和广度持续演进

升级大语言模型

发布30+模型、翻译模型全球第一

升级混元图像3.0模型

混元图像 3.0 在全球26个大模型中，高居第一位，超过nano-banana等顶尖闭源模型

开源多个小模型

0.5B、1.8B、4B、7B等适用于PC、手机、智能驾驶



Tencent Hunyuan

混元3D 3.0 全新发布

3倍建模精度，1536³几何分辨率，36亿建模体素

开源混元世界模型1.1

任何人都能在秒级内从视频或图片创造出专业级3D场景
(10月22日开源)

视频生成模型 混元Video 1.5发布

基于 [DiT] 架构、参数为8.3B的轻量级视频生成模型

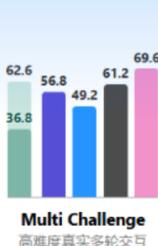
混元推陈出新，不断突破大模型能力天花板

Hunyuan-T1-20250822 HY 2.0-think Qwen3-235B-A22B-2507 (think) DeepSeek-V3.2-Exp (think) Kimi-k2-thinking GPT-5-think (high)

数理逻辑



多轮&指令

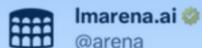


代码&智能体



混元12月最新发 腾讯混元Tencent HY 2.0 Think/Instruct

模型采用混合专家 (MoE) 架构，总参数 406B，激活参数 32B，支持 256K 上下文窗口，推理能力全面提升，综合评测处国内第一梯队。



Vision Leaderboard Shakeup

New model, Hunyuan Vision 1.5 Thinking by @tencent_hunyuan, has entered to tie for #3 in the Vision Arena.

Evaluating AI models with vision adds new complexities when compared to text. To perform well a model must extract information from images, understand it, and combine visual information with text across many use cases such as: diagram interpretations, document parsing, suggesting captions and copy based on visuals, and the ability to solve geometry problems. The list goes on.

Vision combines "seeing/understanding" with reasoning.

Hunyuan Vision 1.5 Thinking is tied for #3 in Vision

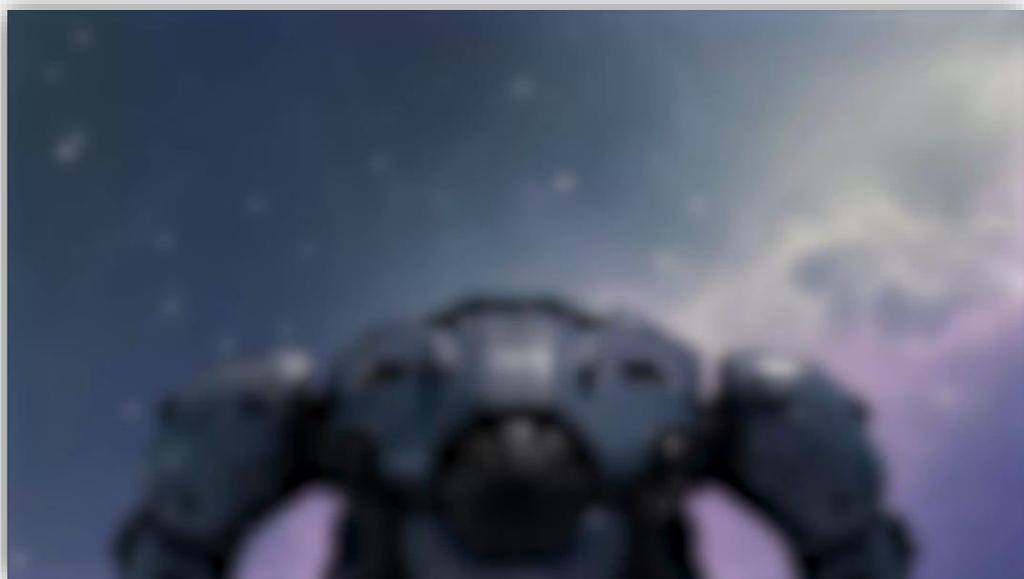
Rank	Model	Score %	95% CI (Std %)	Votes %	Organization %	License %	
1	gemini-2.0-flash	1241	+8	62,824	Google	Proprietary	
1	chatgpt-4o-latest-20250326	1234	+8	14,069	OpenAI	Proprietary	
2	gpt-4.0-gpreview-2025-02-27	1220	+11	2,325	OpenAI	Proprietary	
3	o3-2025-04-16	1213	+8	42,177	OpenAI	Proprietary	
3	gpt-5-chat	1212	+9	35,271	OpenAI	Proprietary	
3	gpt-4.1-2025-04-14	1207	+8	38,606	OpenAI	Proprietary	
3	gemini-2.0-flash	1203	+8	37,468	Google	Proprietary	
3	claude-opus-4-20250314-thinking-10k	1200	+15	1,430	Anthropic	Proprietary	
3	gemini-2.0-flash-preview-09-2025	1201	+16	1,431	Google	Proprietary	
3	hunyuan-vision-1.5-thinking	1200	Proprietary	+14	1,954	Tencent	Proprietary
3	qwen3-v1-235b-a22b-instruct	1197	+15	1,845	Alibaba	Apache 2.0	
5	gpt-5-high	1194	+9	29,364	OpenAI	Proprietary	
5	qwen3-v1-235b-a22b-thinking	1189	+15	1,857	Alibaba	Apache 2.0	
5	claude-3.7-sonnet-20250219-thinking-32k	1189	+15	1,849	Anthropic	Proprietary	

10月LMarena最新榜单 混元视觉模型跻身全球Top 3 国内排名第1

10月8日，国际大模型竞技场LMarena发布最新视觉模型榜单，腾讯混元最新视觉模型Hunyuan-Vision-1.5-Thinking表现抢眼，取得全球第三，国内第一的成绩。

混元的创造能力：混元3D与混元世界模型

混元3D 3.0全新发布



混元世界模型1.1版本 (WorldMirror) 正式发布并开源



混元解锁多模态全新交互体验：视觉理解、视觉推理模型



视觉推理模型 混元 T1 Vision

多模态深度思考
原生长思维链

视觉深度理解
提升5.3%

响应时间更快
速度提升50%



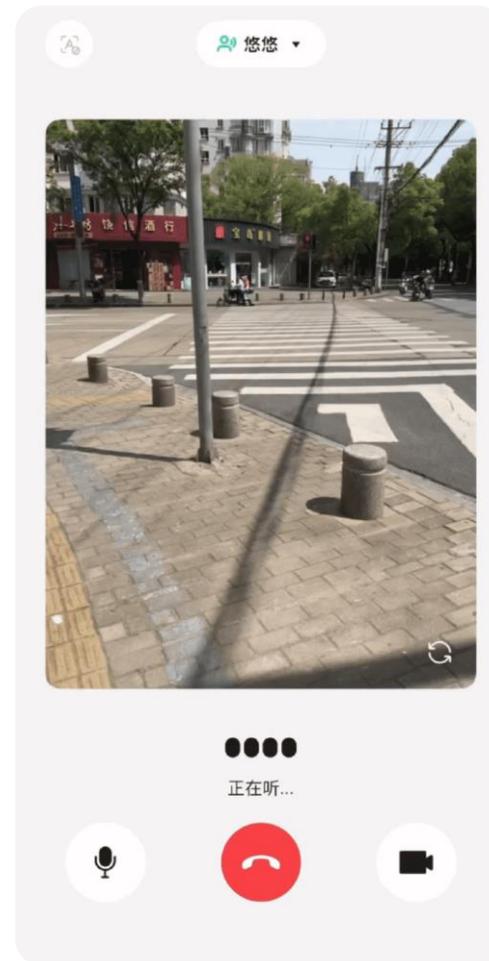
视觉理解模型 混元 Turbo5 Vision

视觉感知理解能力
提升 4.6%

多模态拍照解题能力

数学 提升12%

K12学科 提升 10%



03

AI大模型应用：大模型产品矩阵

——全场景大模型应用产品平台，全方位满足企业业务需求

① 龙虾: OpenClaw 简介

OpenClaw 是什么

1 能动手干活的AI智能体框架

- 是AI 智能体执行引擎，不是单纯聊天工具，能听懂指令并直接操作系统完成任务

2 与元宝等大模型应用的区别

元宝: 回答

① 对话式交互

你问我答，停留在语言层面

② 信息提供者

给你答案，但不执行

③ 被动响应

等待指令，不主动行动

OpenClaw: 执行

① 行动式交互

直接操作工具，完成任务

② 任务执行者

不仅给答案，更帮你做事

③ 主动服务

理解意图，主动完成任务

OpenClaw 像什么

1 数字分身

24小时在线的数字版的你替你处理日常事务。

- 知道你的偏好
- 记住你的习惯
- 越用越懂你

2 私人秘书

专业的行政助理处理邮件、日程、会议。

- 高效专业
- 主动服务
- 随时待命

3 自动化管家

智能的生活管家自动化处理日常琐事。

- 流程自动化
- 智能提醒
- 生活优化

OpenClaw 关键能力

1 管理工具

- 邮箱管理: 自动分类、回复、跟进
- 日程协调: 智能安排、冲突检测
- 任务追踪: 待办清单、进度提醒

2 自动 workflow

流程自动化: 重复任务一键执行

多工具联动: 跨平台无缝协作

智能触发: 条件判断、自动响应

3 多平台交互

聊天应用: 企微、QQ、飞书、钉钉等

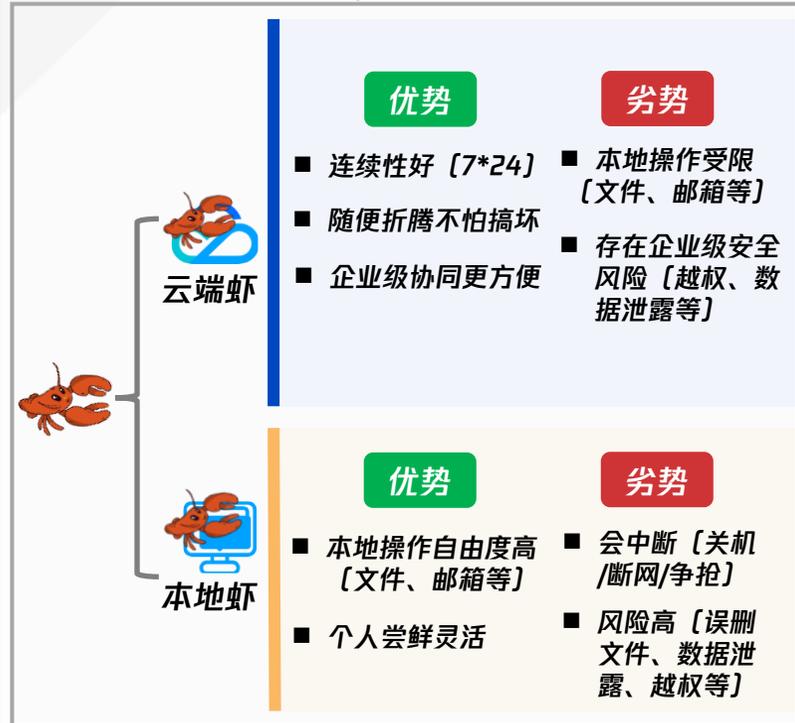
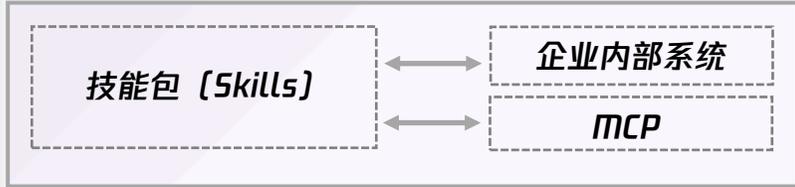
系统级访问: 文件、命令、脚本

浏览器控制: 网页操作、数据提取

安全与分险: 隐私安全, AI误操作, 执行边界

腾讯云安心“养虾”解决方案产品矩阵

安心“养虾”需求



MaaS API

腾讯云安心“养虾”解决方案产品矩阵

SkillHub ☆官方认证 加速下载 安全审计
 专为中国用户优化的AI Skills社区，精选推荐，高速下载体验，轻松查找ClawHub 1.3万个Skills

ClawPro [企业版] 企业
 适合各种规模的企业使用，具有统一管控、安全性高、运维便捷等优势

轻量云OpenClaw 企业 个人
 适合个人或者小规模企业使用，具有成本低、一键部署、Web配置等优势

云桌面OpenClaw 企业 个人
 适合个人或者小规模企业使用，具有成本低、一键部署、熟悉的windows视窗操作界面等优势

WorkBuddy 企业 个人
 腾讯纯自研产品，与CodeBuddy同源，更新迭代快，架构优秀、性能好、能力强，稳定性高等

qClaw 个人
 腾讯电脑管家出品，免费下载安装，开箱即用，微信直连远控，适合个人尝鲜

AI Agent MaaS与异构算力方案

大模型API
 [混元、DS、Kimi、GLM、Minimax等模型最新版]

高性能应用服务HAI 国产GPU
 [定制模型推理服务] [国产化、自研推理需求]

Codebuddy
 代码助手，支持自研Skill
 知识库：ADP、乐享
 网络：Github/Clawdhub加速

AI Agent安全方案

AI Agent安全中心
 资产识别、Skills风险扫描，敏感行为识别、数据泄露检测、破坏文件拦截、安全审计等

MCP安全网关
 用户channels和访问认证、凭据管理、提示测攻击、web攻击检测和拦截

NDR流量安全检测
 网络层流量审计和溯源

.....

iOA AI-Agent安全
 Agent准入管控、运行时监控、高危Skill入侵防护、数据外泄拦截、审计与溯源等

龙虾管家 端免费
 为本地运行的AI Agent提供一站式安全防护，覆盖系统、Skill、支付、Prompt安全等

AI Agent算力和支撑方案

云服务器CVM
 性能强劲，适合大规模部署

轻量云服务器Lighthouse
 碎片资源，低成本资源套餐

云桌面
 支持windows/Ubuntu系统

对象存储 - Vector bucket
 长期记忆降本增效方案

沙箱AGS
 安全隔离、极致性能、浏览器/代码多种形态

腾讯OpenClaw企业版打造企业专属龙虾



管控端界面



员工端界面

员工端：一键启用OpenClaw

一键创建

无需关注底层基础设施配置，即刻生成专属助理

可视化配置

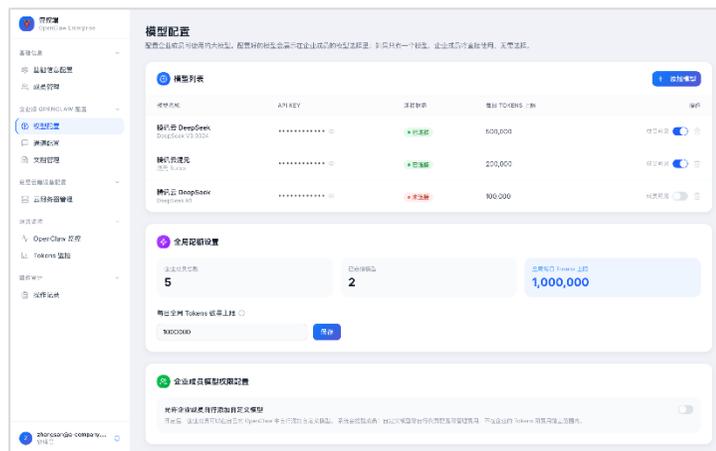
Web可视化配置通道，无需登录服务器，简单便捷

开始使用

无缝接入主流IM工具，随时随地开启智能对话

配额追踪

模型配额实时预览，掌握使用情况



管控端：企业级管理

定制镜像

上传企业专属OpenClaw应用镜像，定制化部署

可视化配置

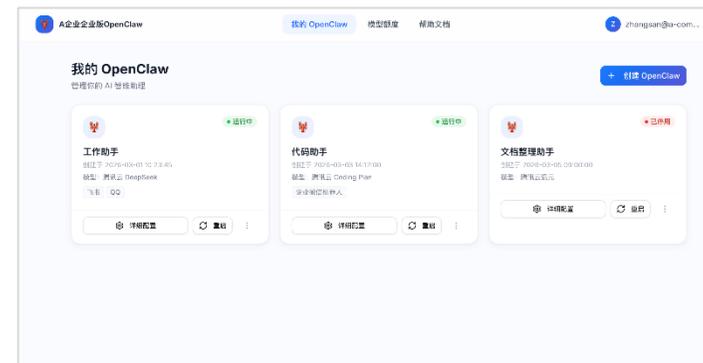
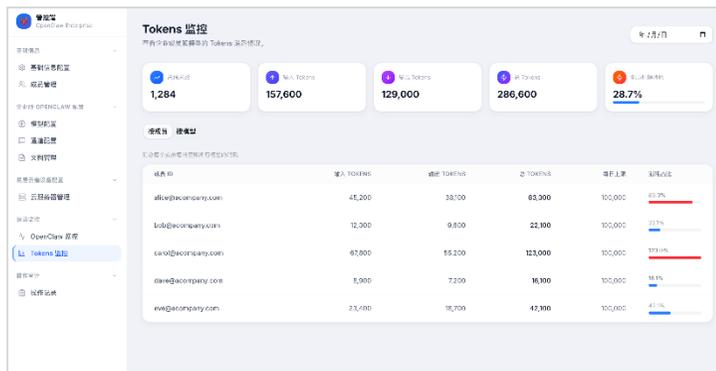
按需配置模型参数，一键开启主流IM通道

配额管理

支持按员工、按总量等多维度配额管理，灵活分配

用量监控

OpenClaw调用数、Token消耗数实时监控，数据透明



一键获取专属管理门户，彰显企业形象
百元起开启 AI 之旅，按需计费灵活配置

产品定位

企业一键开通，员工人人拥有专属 OpenClaw 数字助手



产品交付物

- 管控台：员工端 + 管控端
- 基础算力：计算+存储+网络等

[🔗：腾讯云OpenClaw企业版·产品介绍文档](#)

为啥选择腾讯云OpenClaw企业版

Lighthouse成熟经验

腾讯轻量云百万级OpenClaw用户
同源同构，同款应用镜像

高效skill安装通道

专为中国用户优化的 AI Skills 社区，
腾讯国内首发

四大IM首发适配

QQ、企微、钉钉、飞书，QQ/飞书
扫码即接入

网络隔离与端口加固

多租户虚拟机强隔离，随机端口自
动适配强化安全

独家的配置面板体验

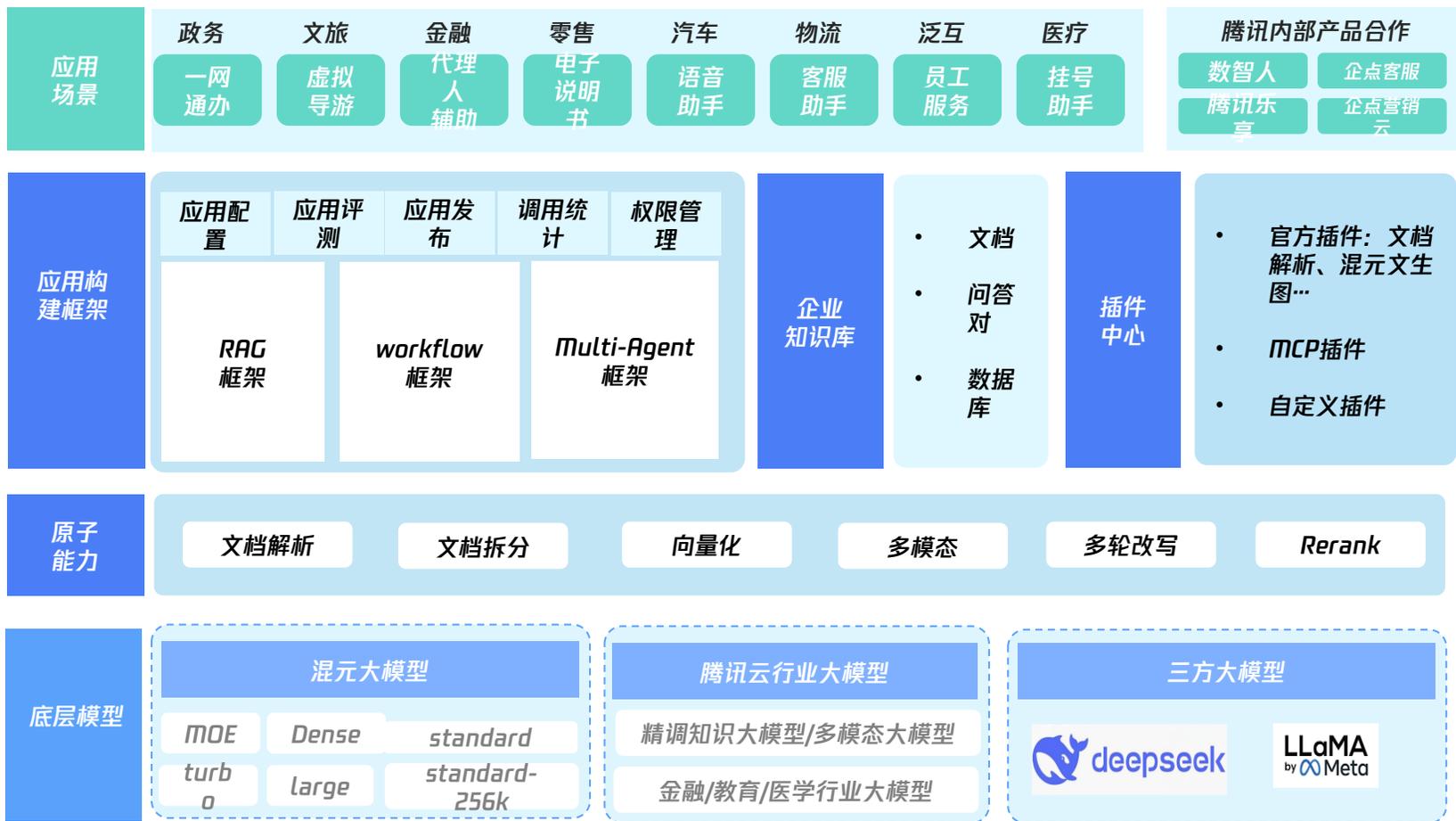
独家TAT、Orcaterm、AI助手功能，
贴近应用

CVM定制AI机型Ai2

面向AI Agent场景全新机型，刊例
价为标准型40%普惠定价

② 智能体开发平台：提供RAG、Workflow等多种开发框架，助力企业AI落地

腾讯云智能体开发平台，是面向基于大模型的应用构建平台。能够结合企业专属数据，助力企业更快更高效地搭建Agent、RAG、工作流等多种应用框架，快速部署符合个性化业务需求的大模型应用。



RAG效果保持领先

- 知识问答部分领域效果领先：在图文混排文档问答、复杂大表问答等领域相比友商有一定优势。

具备multi-agent、工作流复杂应用构建能力

- Multi-agent模式发布，主Agent和子Agent协同响应对话，支持官方/自定义插件调用，工作流Agent对话灵活性表现好。
- 新版工作流发布，支持14个常用画布节点，支持客户通过拖拽大模型节点、参数提取节点、知识问答节点等常用原子能力编排业务流程。

自研+开源模型驱动，各取所长

- 自研模型：混元全系列模型；自研精调知识在无关知识拒答、多模态阅读理解上有优势；在医学、教育等行业大模型上有落地标杆。
- 三方开源模型：接入DeepSeek R1/V3,叠加联网搜索+RAG能力，进一步放大RAG优势。

落地案例：某头部车企的智能客服助手

项目背景

- **客户预期：**通过大模型能力的加持，提高C端咨询的机器人独立解决率，较少人工接待会话量，节省客服中心人力成本
- **使用场景：**用户的信息咨询，如车辆使用疑问、日常故障处理、保修保养政策等
- **面向人群：**车主

项目方案

1. 集成腾讯云智能体开发平台能力，接入到小程序/App/官网/公众号等多个渠道支持车主咨询，提升服务的准确度和效率；
2. 基于腾讯云智能体开发平台能力提炼历史客服知识库信息，作为企业知识库的有效补充，进一步丰富专业客服知识体系，降低知识库的整理工作量。

方案效果

- **智能客服机器人独立解决率已从37%提升为84%**，高效解决了车企售后场景问题咨询，有效减少人力成本的投入

产品示例

我是  AI智能客服，您可以直接向
我提出您的问题。我将竭诚为您服务。请
您注意：相关回复内容由大模型系统根据
企业知识库生成，不具备解决相关问题、
纠纷的法律效力，仅供您参考之用。如您
有任何疑问，请您转人工为您服务

 重置发动机机油保养数据 

您好！关于  重置发动机机油保养数
据的操作，不同车型和年份的  可能
有细微的差异。例如，2021年  版
本和2023年  版
本的操作步骤略有不同。请问您是想了
解哪个年份和型号的  的重置步骤
呢？这样我可以为您提供更准确的指导。

2023年  版本 

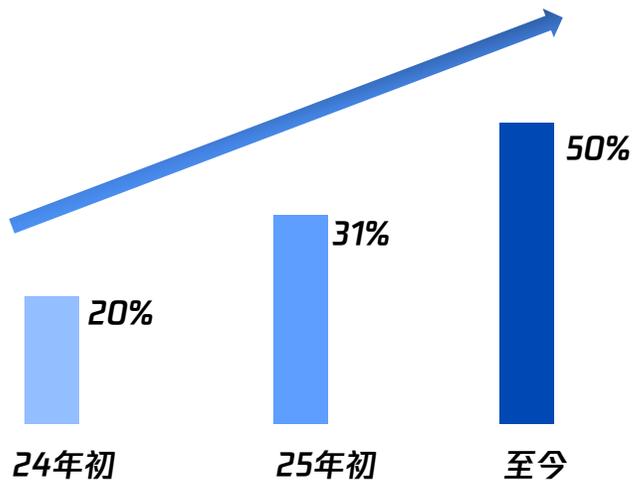
2023年  版本
重置发动机机油保养数据的操作步骤如
下：

1. 选择多信息显示屏上的 
2. 按下仪表控制开关选择“车辆设定”，
然后按住OK。
3. 按下仪表控制开关选择“机油保养”，
然后按下OK。
4. 按下仪表控制开关选择“是”，然后按
下OK。

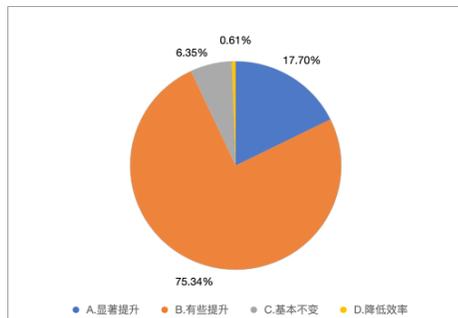
完成重置程序时，多信息显示屏上将显示
信息1。

③ 腾讯研效率工具平台: CodeBuddy

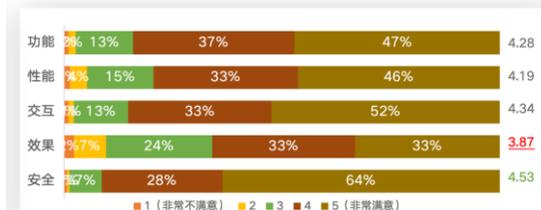
作为腾讯研发智能化的第一批产品
CodeBuddy提供AI辅助编码和AI程序员能力。
随着产品能力的发展, CodeBuddy在集团内部快速的被接受, 使用, 和依赖



代码生成率50%



满意度: 4.54 / 5



93%的受访使用用户认为AI助手可以帮助自己提效



好评率 4.54 / 5

82.26%

整体活跃率 (5W+T族总用户)

> 50%
代码生成率

> 28%
补全采纳率

89%

标杆BG活跃率

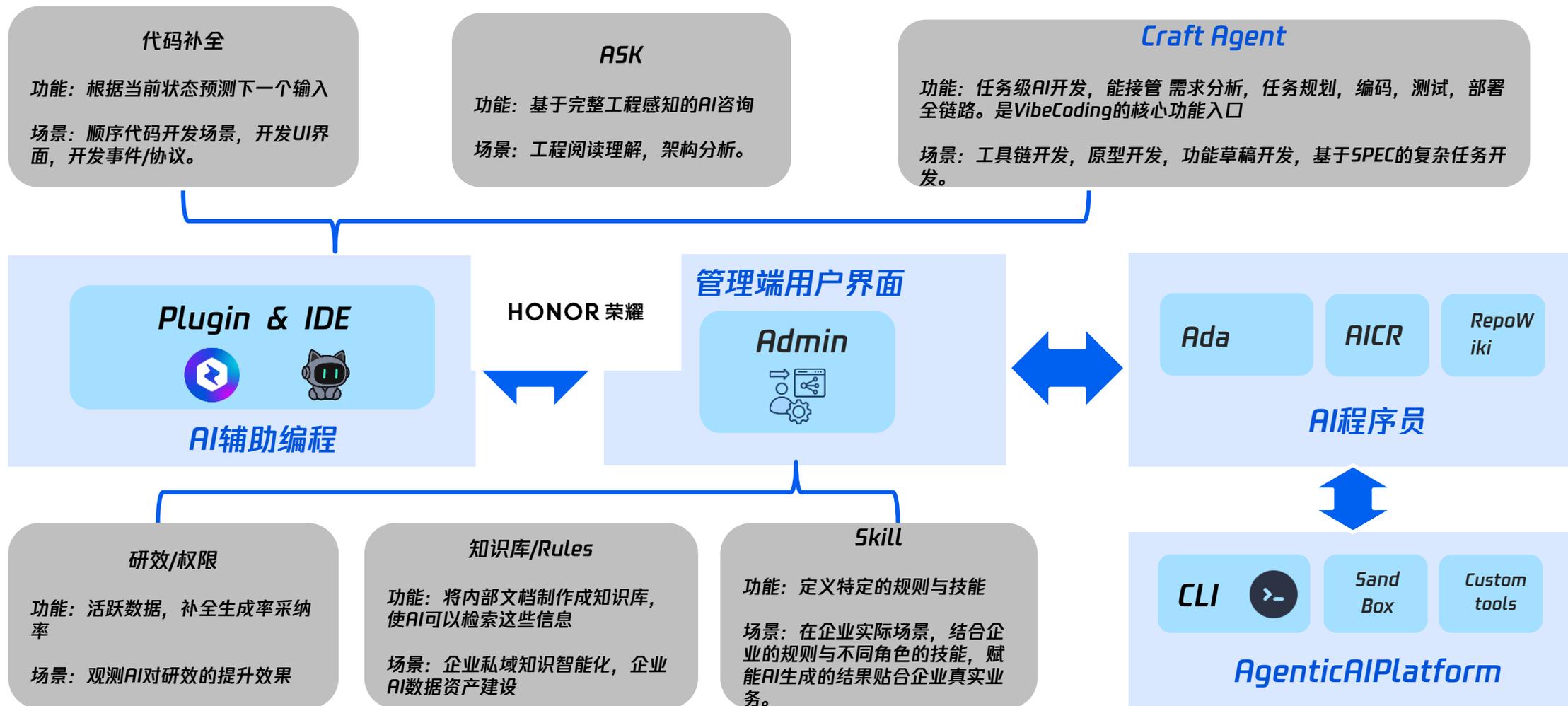
52.9%
代码生成率

30%
补全采纳率



是腾讯80%工程师的选择

CodeBuddy 产品核心功能架构



HONOR 荣耀



小米集团

项目背景

小米集团始终致力于提升研发效能与质量，并持续关注AISE行业进展。随着Copilot在全球获得广泛认可，小米于2023年底启动AI辅助编码的探索，手机软件部分别组建团队推进IDE插件和代码大模型的训练与微调。经过四个月建设，虽取得初步进展，但在代码采纳率、生成效果及IDE体验等方面仍未达到内部预期，且存在代码规范不一、模型微调效果不佳等问题。因此，小米决定引入成熟的商业化解决方案。

客户痛点/需求:

1. 插件研发：人力投入较大，建设效果慢，各类IDE的兼容，市场上插件功能的追赶，开源社区活跃度低；
2. 模型微调：开源模型微调困难，结果不收敛，并且存在大量的bad case无法解决。
3. 补全体验：自研IDE插件体验较差，补全响应慢、截断不准确
4. 工程分析：自研的IDE插件，在跨文件检索、匹配、推理方面存在技术瓶颈，对于一些复杂的业务代码，补全效果不理想



解决方案

1. IDE 插件融合: 使用腾讯云 AI 代码助手插件内核，将插件的样式、logo、登陆、反馈、升级机制等内容保持和小米自研IDE插件一致。
2. 模型训练: 基于腾讯代码大模型，结合小米手机的业务代码进行了增强训练，训练之后，手机操作系统团队的代码补全生成率**↑16%**。
3. 联合运营: 配合小米MIT生产力团队，对小米的研发人员进行了多次 AI 编码最佳实践培训，提升用户活跃度



项目收益

内部用户: **10,000+**人

代码采纳率: **27%**

代码生成率: **36%** [等效释放 3600+人月]

插件崩溃率: **< 0.1%**

整体研发提效: **20%**

④ 通过TRTC端到端集成，快捷实现AI大模型对话互动场景



客户端和服务端的一体化集成方案，提供智能打断、降噪能力。

功能特点

实现超低延迟的AI对话

音视频端到端延迟低于300ms，AI对话延迟低于1000ms。

精准的ASR识别

支持多种语言，包括英语、西班牙语、日语、韩语、中文以及23种方言和130种国际语言。

TTS无缝集成

支持超过53种音色，合成音色接近真人，拟真度，支持纯中文、纯英文、中英文混合、粤语内容

端到端完美集成

集成ASR、TTS、智能体/大模型平台等腾讯优势产品能力，提升开发效率，增强产品稳定性。

技术优势

双讲打断

AI聊天过程，实时打断，真正实现与人聊天的真实体验。

AI降噪

全新降噪引擎，输入端音频实时AI降噪，提升ASR的识别准确度。

服务稳定

全球2800+节点，多服务容灾，高稳定超低延迟。

基于大模型能力的音视频产品TWeTalk充分满足AI玩具需求

——腾讯云音视频TWeTalk是面向智能硬件的AI对话软硬一体化产品方案，TWeTalk与主流嵌入式芯片及模组厂商合作，整合端侧、云端与通信能力。该方案已应用于AI陪伴玩具、机器人、智能穿戴（手表、眼镜）、耳机同传、智能点餐、导览、AI面试等AIoT场景，实现人与硬件的自然“对话”（Talk）。



语言端到端与音色定制：独特的意图理解ASR引擎，不仅支持语音识别，更能精准捕捉用户的情绪与意图；TTS支持音色定制；精调大模型适配玩具、机器人等场景，提供准确执行function call的高性价比引擎。



情绪识别：依托先进的语音处理技术与精调场景大模型，精准捕捉语调、语速、语义的细微特征以感知用户情绪，及时给予礼貌、贴合心境的回应。



腾讯生态产品集成：整体集成腾讯微信通话、支付、QQ音乐等腾讯系产品，打造AI玩具亮点，为产品增加附加值与用户吸引力。



汤姆猫 AI 业务联创韩承晏表示：“腾讯云TWeTalk全球化方案的本地化部署与语言技术能力，为我们在北美市场的规模化落地提供了关键支持。双方合作从产研到市场落地无缝衔接，我们高度看好未来合作，期待共同拓展全球AI陪伴版图。”



以我所能 为你而+