

AI 大模型创业格局报告

序 言

2022年11月ChatGPT发布，引爆国内百模大战，创业玩家如雨后春笋，互联网巨头也纷纷下场，巅峰时期国内公开大模型数量超越300个。

如今一年半过去，巨头们依托固有业务，更早形成壁垒，率先辍。而创业玩家这边，胜局也已锁定，大模型创业第一梯队已经形成，目前或许只有少数几家公司拿到了和巨头角逐下半场的门票

他们分别是创业六小强——智谱 AI、MiniMax、阶跃星、百川智能、月之暗潮零一万物，这六家公司都转起了一个关键飞轮——模型、应用和融资飞轮，这使得他们能够与巨头掰腕子。另外，这一赛道上还有两位值得特别关注的玩家——疆智能和 DeepSeek。

他们共同组成了国内大模型创业玩家“6+2”格局。许多人认为，继移动互联网之后的 AI 时代，新的巨头或将从这八家公司中诞生。

而放眼全球，美国大模型行业同样存在类似的阶段性竞争格局变化，像曾经的移动互联网时代一样，中美再次在两个“平行宇宙”中上演新的历史进程……

在本篇报告中，量子位智库将以一个逐步放大的视角，从国内创业玩家“6+2”格局、国内巨头玩家格局、海外玩家格局三个层次，力求为读者描绘大模型行业格局完整画卷。

目 录

1.0 国内创业玩家“6+2”格局

2.0 国内巨头玩家格局

3.0 海外玩家格局

4.0 新阶段竞速标准

ights

1.0

国内创业玩家“6+2”格局

insights

国内创业玩家“6+2”格局总览

| 创业六小强 | 分类 | 创始人 | 模型 | C 端产品 | B 端产品 | 最新估值 |
|-------|-----|-----|----------|--------------|---------------------------------------|-----------------|
| | | 学院派 | 唐杰 | GLM | 智谱清言、清影 | API、云端私有化、本地私有化 |
| | 青年派 | 闫俊杰 | abab | 星野、海螺、Talkie | API | 约 25 亿美元 |
| | | 杨植麟 | Moonshot | Kimi | API | 约 30 亿美元 |
| | 大佬派 | 姜大昕 | Step | 跃问、冒泡鸭 | AI 金融、AI 内容创作 | 约 20 亿美元* |
| | | 王小川 | Baichuan | 百小应、健康顾问 | MaaS+AaaS 行业解决方案 | 超 20 亿美元 |
| | | 李开复 | Yi | - | 数字人解决方案 全行业 ToB 解决方案、AI Infra 解决方案 | 超 10 亿美元 |

*为撮估值

6
+
2

创业特别关注

| 模型 | 创始人 |
|----------|------------|
| | |
| MiniCPM | 李大海 刘知远 |
| | |
| DeepSeek | 梁文锋 |

2019.06 智谱 AI 成立

2020.06.11 OpenAI 发布 GPT-3

2021.12 MiniMax 成立

2022.08 面壁智能成立

2022.11 OpenAI 发布 ChatGPT, 引爆百模大战

2023.03 月之暗成成立

2023.04.06 阶跃星辰成立

2023.04.10 百川智能成立

2023.05.16 零一万物成立

2023.07.17 深度求索成立

国内创业玩家核心竞争力四维度

竞争力维度

具体内容

模型

在模型维度上，本章将根据各家创业公司特点有侧重地介绍其研发历程、模型能力特点、架构特色、矩阵构成、Benchmark 成绩、AGI 路径规划等情况中的一到三项。

应用

在应用维度上，本章将根据各家创业公司特点有侧重地介绍其大模型应用矩阵、特色应用、主力应用、实现功能、用户数据、战略特点、生态布局、付费模式等情况中的一到三项。

融资

在融资维度上，本章将介绍各家创业公司融资历程，具体包括历次融资时间、轮次、金额、投资机构。

团队

在团队维度上，本章将介绍各家创业公司核心团队成员情况，具体包括教育背景、学术影响力、重要衔、相关工作履历等信息。

ights

1.1

创业六小强玩家

insights

1.1.1

智谱 AI:
中国版 OpenAI

1.1.4

阶跃星辰

1.1.2

MiniMax

1.1.5

百川智能

1.1.3

月之暗面

1.1.6

零一万物

智谱 AI——模型：清华系成果转化，自研底层架构，对标 OpenAI

脱胎于高校实验室，技术成熟自然转化：智谱 AI 脱胎于清华 KEG 实验室，是清华计算机系技术成熟到一定程度的转化成果，于 2019 年 6 月正式成立，2022 年 8 月发布自研百亿参数大模型 GLM，彼时 GPT-3.5 还未发布。

自研底层框架打破垄断，矩阵全面对标 OpenAI：2022 年 3 月，唐杰团队提出了全新自研的通用预训练框架 GLM，打破当时西方垄断主流大模型预训练框架技术路径的局面。智谱 AI 在模型和产矩阵甚至投策略上均对标 OpenAI，最新的 GLM-4 基座大模型能力多方位接近世界领先水平。



OpenAI vs 智谱

| | | |
|----------------------------|------------|-----------------|
| ChatGPT | 应用产品 | ChatGLM 智谱清言 |
| GPT | 底层框架 | GLM |
| DALL·E | 图像生成 | CogView |
| Sora | 视频生成 | CogVideo 清影 |
| Codex | 代码 | CodeGeeX |
| GPT-4o | 多模态 | CogVLM |
| platform.openai | MaaS 平台 | bigmodel.ai |
| gptstore | Agent 开发平台 | 智谱清言智能体中心 |
| Helion、Rain AI、Anthropic 等 | 产业投资 | 面壁智能、生数科技、无问芯穹等 |

GLM-4 模型主流 Benchmark 成绩

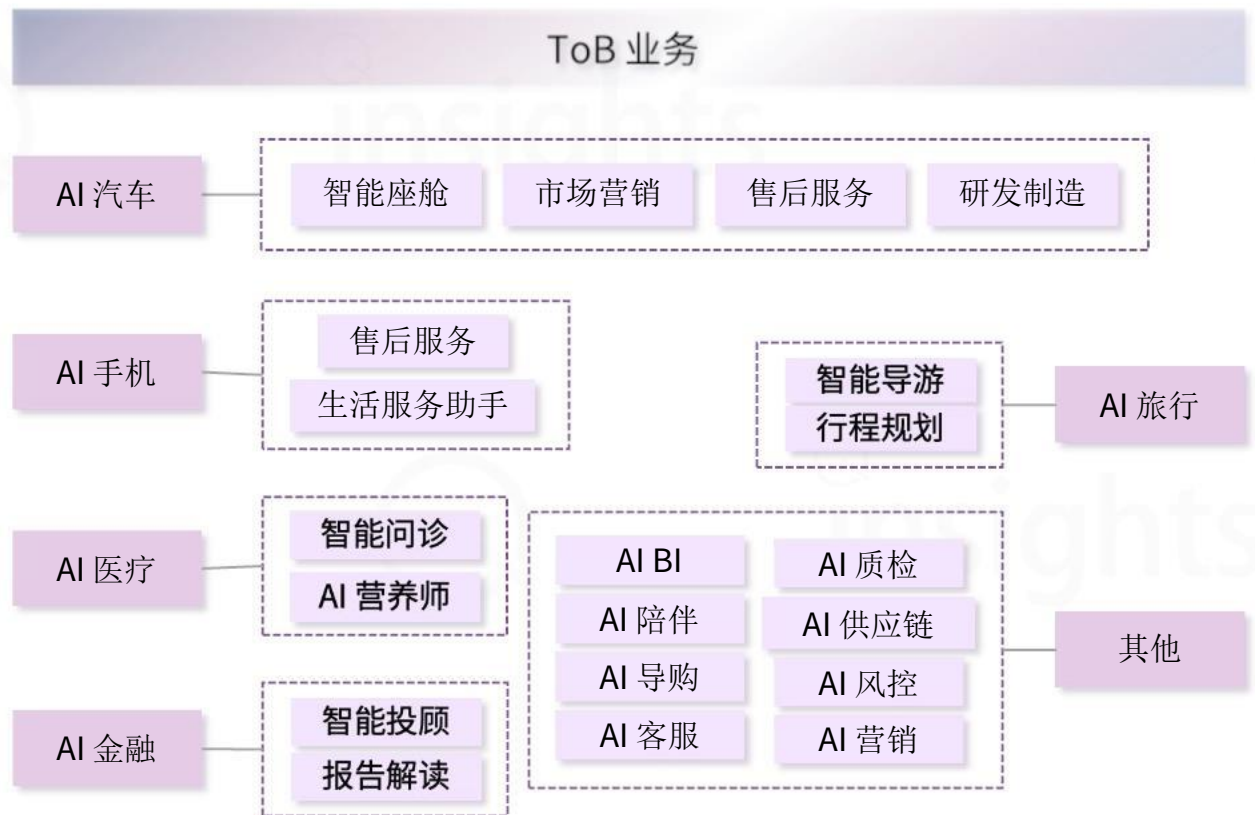
| Model | MMLU | GSM8K | MATH | BBH | GPQA | HumanEval |
|--------------------------|------|-------|------|------|------|-----------|
| GPT-4 (0314) | 86.4 | 92.0 | 52.9 | 83.1 | 35.7 | 67.0 |
| GPT-4 Turbo (1106) | 84.7 | 95.7 | 64.3 | 88.3 | 42.5 | 83.7 |
| GPT-4 Turbo (2024-04-09) | 86.7 | 95.6 | 73.4 | 88.2 | 49.3 | 88.2 |
| Claude 3 Opus | 86.8 | 95.0 | 60.1 | 86.8 | 50.4 | 84.9 |
| Gemini 1.5 Pro | 85.9 | 90.8 | 67.7 | 89.2 | 46.2 | 84.1 |
| GLM-4-9B-Chat | 72.4 | 79.6 | 50.6 | 76.3 | 28.8 | 71.8 |
| GLM-4-Air (0605) | 81.9 | 90.9 | 57.9 | 80.4 | 38.4 | 75.7 |
| GLM-4 (0116) | 81.5 | 87.6 | 47.9 | 82.3 | 35.7 | 72.0 |
| GLM-4 (0520) | 83.3 | 93.3 | 61.3 | 84.7 | 39.9 | 78.5 |

智谱 GLM-4(0520) 在 GSM8K、MATH、BBH、GPQA、HumanEval 等多个国际主流 Benchmark 上成绩超过 GPT-4(0314)。

智谱 AI——应用：工程圈，侧重 ToB，商业化进程身位领先

ToB圈突出，主打 MaaS 平台：智谱脱胎的清华 KEG 知识工程研究室自 90 年代就强调工程落地，造就了团队将技术落地到需求中的做事风格。智谱团队以 ToB 背景为主，商业化重心在 B 端，智谱大模型开放平台目前已有超 40 万的企业用户，主覆盖 金融、教育、互联网、零售、汽车、能源传统制造业等行业。

ToC 侧重生产力应用，视频生成功能“清影”上线：智谱 ToC 业务目前仍以免费为主，主力应用“智谱清言”，截至 2024 年 6 月历史下载量突破 500 万，同时智谱清言也赋能体开发平台，目前已有超过 30 万个智能体活跃在智谱清言 APP 可供使用，2024 年 7 月新上线视频生成功能“清影”，对标 Sora，免费公开使用不限次。



智谱清影 AI 生成视频功能网页端

依据量子位智库发布的《AI 智能助手 APP 8 月榜单》，智谱的 ToC 主力应用智谱清言下载量突破 500 万，位列第七。（完整报告数据请查看量子位智库公众号）

API 单日调用量
600 万+

API 单日调用 tokens
400 亿+

服务企业客户
30 万+

智谱 AI——融资：融资轮次多、估值先，全产业链投积极布局生态

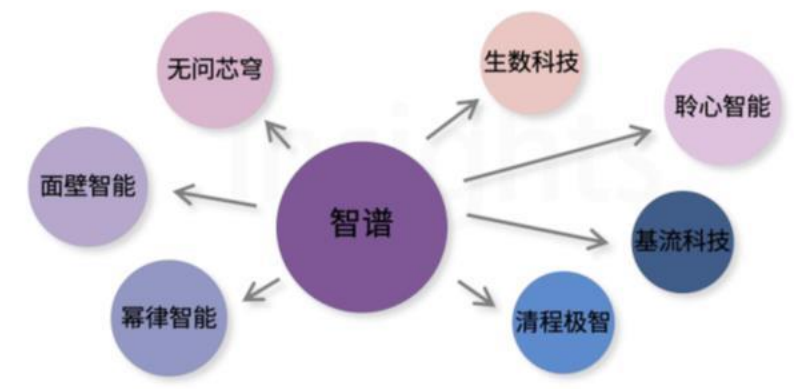
融资金额及估值处于第一梯队：智谱是国内估值最高的大模型初创公司之一，最新一轮融资估值据传为 30 亿美元，从团队背景、模型开发进展、商业化方向及资金储备情况来看公司处于国内 AI 创业公司头部。

积极布局产业生态，投多家技术优势公司：为打造自己的 AI 大模型生态圈，智谱参与了十多家 AI 大模型产业相关创业公司的投，被投企业多为清华系，且以技术导向为主。

智谱融资历程

| 时间 | 轮次 | 金额 | 估值 | 投资机构 |
|---------|------|---------|----------|---|
| 2024/09 | B+ 轮 | 数十亿 | 200 亿 | 中关村科学城 |
| 2024/05 | B+ 轮 | 约 4 亿美元 | 约 30 亿美元 | Prosperity7 Ventures |
| 2024/03 | B+ 轮 | - | - | 顺禧鑫 |
| 2024/01 | B+ 轮 | - | - | 光速光合 |
| 2023/10 | B+ 轮 | - | - | 蚂蚁集团 |
| 2023/09 | B+ 轮 | 12 亿 | 25 亿 | 好未来、小米、红杉、美团、腾讯投资、金山软件、阿里巴巴、顺为资本、高瓴资本、君联资本 |
| 2023/07 | B+ 轮 | 约 3 亿 | 40 亿 | 美团 |
| 2023/05 | B+ 轮 | 约 4 亿 | 40 亿 | 今日资本，蚂蚁金服，光速中国 |
| 2022/09 | B 轮 | 约 2 亿 | 20 亿 | 启明创投，君联资本，华控鑫 |
| 2021/09 | A 轮 | 约 1.5 亿 | - | 图灵创投凌云光、华控鑫、将门创投海南智桥、IMO Ventures、达晨财智、北京荣品 |
| 2021/04 | 天使轮 | - | - | 北京创新智源科技有限公司 |
| 2019/08 | 种子轮 | - | - | 中科创星华控技术转移有限公司、清控资管集团 |

智谱投资布局



智谱产业生态布局

对创业者

发起总额为 10 亿人民币的大模型创业基金，支持大模型原始创新，覆盖 大模型基础、底层算、芯片优化、行业大模型和超级应用等各方面。

对学术界

分别联合 CCF 中国计算机学会、中国中文信息学会社交媒体处理专委会发起了 CCF- 智谱大模型鑫、SMP- 智谱大模型交叉学科鑫，用于支持大模型理论、基础、模型、

应用等相关研究和跨领域交叉创新。

对开源区

成立大模型开源鑫 1000 张 GPU、1000 万人民币和 1000 亿 tokens 来支持人工智能开鑫

智谱 AI——团队：学院派代表，清华计算系人才众多

依托清华 AI 人才资源：截至 2024 年 9 月，公司员工共计 600 余人，其中大模型研发人员接近半数，其余为客户服务落地人员及产品人员。公司借助清华系高校资源，持续吸引高端人才。

团队具备从研发到落地全栈能力：公司团队源自清华知识工程研究室 KEG，该实验室自 90 年代成立起就走的是“科研+工程化落地”路线，团队成员具备深厚技术积累及综合运营经验；创始人唐杰教授是人工智能领域顶尖专家，具备 20 余年人工智能领域研究经验。



唐杰
创始人
技术顾问



- 清华大学计算机系本科、硕博、中国 IEEE Fellow 标准化分会委员
- 清华大学计算机系教授，研究方向知识图谱
- 曾主导开发全球最大万亿级超大规模预训练模型悟道



张鹏
CEO

- 电子学会标准化工作委员会机器人技术委员会委员
- 在 ICML、ISWC 等顶级会议上发表 10 余篇文章



张帆
COO

机硕士、法国梅斯大学计算机硕士、巴黎第十一大学计算

- 曾任大搜车集团 CTO，曾创立元因智能任 CEO，创立妙计旅行任 CEO，智能技术领域连续创业者
- 曾在字节，腾讯，搜狗从事技术管理工作



王绍兰
总裁



- 曾任清华数据科学研究院科技大数据研究中心副主任
- 凹凸科技北京公司总经理
- 获得授权发明专利 15 项，美国发明专利 2 项



刘德兵
董



- 20 余年技术和政府项目综合经验
- 清华数据科学研究院科技大数据研究中心副主任
- 中科院计算所博士，师从高文院士
- 获得授权发明专利 40 余项，国际标准一项

1.1.1 智谱 AI

1.1.4 阶跃星辰

1.1.2 MiniMax:
打造爆款 AI 内容平台

1.1.5 百川智能

1.1.3 月之暗面

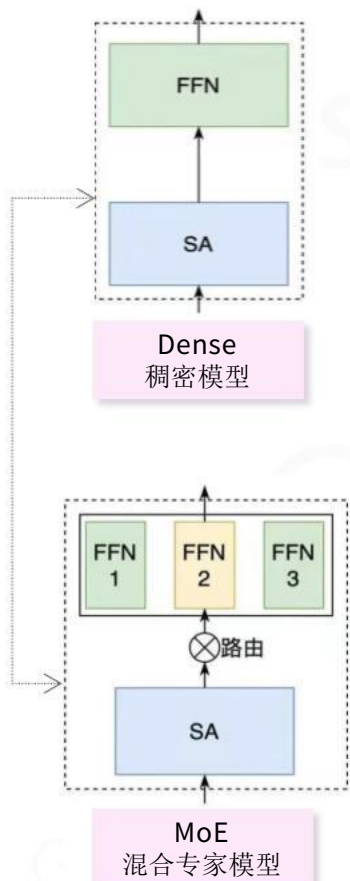
1.1.6 零一万物

MiniMax——模型：选择混合专家架构技术路径，打造多模态模型

国内最早研发 MoE 及 Linear Attention：2023 年 6 月，MiniMax 在 MoE 混合专家架构仍是行业非共识时选择 All in MoE，彼时行业大部分公司仍在训练 Dense 稠密模型。2024 年 1 月，MiniMax 发布千亿参数 MoE 架构模型；4 月底，推出万亿参数 MoE 架构模型。即将发布的 abab 7 模型将基 MOE+ Linear Attention。

围绕产需求打造全矩阵多模态模型：截至 2024 年 9 月，MiniMax 已推出视频大模型、音乐大模型及语音大模型，最新主力模型 abab 6.5 在多项核心能力上开始接近世界最领先大语言模型水平。

模型架构对比



- 由自注意力机制层（SA）和前馈神经网络层（FFN）组成，每一层都与前层直接相连。
- 每次前向传播中所有的参数都被激活和使用，需要的计算量和显存更大。

- 自注意力机制层（SA）后接了多个前馈神经网络层（FFN）。
- 在每次前向传播中，由路由机制选择性激活部分 FFN。
- 可减少计算量和显存需求，同时提高模型在同等条件下的灵活性和容量。

模型矩阵



单日交互次数
30 亿次

日均处理
Tokens
3 万亿

主流测试集成绩

| | 英文综合 | | 中文综合 | 知识 | 基础算数 | 数学解题 | 逻辑推理 | 编程 | 指令遵从 |
|---------------------|----------|------------|------|-------|------|------|-----------|--------|------|
| | MT-Bench | AlignBench | MMLU | GSM8K | MATH | BBH | HumanEval | IFEval | |
| abab-6.5-2024 0415 | 8.77 | 7.42 | 78.7 | 91.7 | 51.3 | 83.1 | 78.0 | 78.9 | |
| abab-6.5s-2024 0415 | 8.69 | 7.18 | 73.0 | 87.5 | 42.6 | 76.8 | 68.3 | 72.8 | |
| GPT-4 | 8.96 | 7.53 | 86.4 | 92.0 | 52.9 | 83.1 | 67.0 | 85.6 | |
| GPT-3.5 | 8.21 | 6.08 | 70.0 | 57.1 | 34.1 | 66.6 | 48.1 | 70.5 | |
| Gemini 1.5 Pro | 8.93* | 7.33* | 81.9 | 91.7 | 58.5 | 84.0 | 71.9 | 82.3* | |
| Claude 3 Opus | 9.00* | 7.62* | 86.8 | 95.0 | 61.0 | 86.8 | 84.9 | 91.9* | |
| Claude 3 Sonnet | 8.47* | 6.70* | 79.0 | 92.3 | 40.5 | 82.9 | 73.0 | 87.8* | |
| Claude 3 Haiku | 8.39* | 6.42* | 75.2 | 88.9 | 40.9 | 73.7 | 75.9 | 85.4* | |

abab 6.5 与世界领先语言模型在多个开源测试集上的成绩对比

abab 6.5 系列包含两个模型：abab 6.5 和 abab 6.5s。在各类核心能力测试中，abab 6.5 开始接近 GPT-4、Claude-3、Gemini-1.5 等世界上最领先的大语言模型。

MiniMax——应用：以娱乐陪伴方向为主，专注打造超级应用

ToC 打造超级应用工厂：MiniMax 是国内创业公司里做应用产最早、最多的公司，2022 年早于 ChatGPT 一个月上线 Glow，又陆推出了 AI 内容平台应用暨、Talkie，以及生产力助手海螺 AI。8 月最新发布海螺 AI 创意视频平台，可调用最新视频生成模型，视频生成能力尤其是物表情生成方到海内外好评。

ToB 布局开放平台：ToB 方向布局 API 开放平台，目前已接入 3 万余家企业客户，主地场景包含零售、教育、办公、智能硬件、招聘、客服、聊天等。

国内外 AI 内容平台应用



国内暨：暨于 2023 年 9 月上线，用户可以在暨上通过简单描述定制 AI 智能体的形象、声音、人设、技能，并在后续对话中不断调整强化。也可以选择首页推荐的数百万款由用户创建的智能体进行多模态沉浸式互动。

此外，星野还探索了订阅制和抽卡制的付费模式、以宝石为主的经济体系和以星念为中心的用户共创激励体系。

依据量子位智库发布的《AI 陪伴 8 月 APP 榜单》，星野 APP 下载总量超 1400 万，月新增下载及 DAU 均位列国内同类 AI 产第一。



海外 Talkie：Talkie 于 2023 年 6 月上线，主要面向北美市场。产品玩法与星野类似，用户可以定制 AI 智能体的形象、声音、人设、技能，并进行卡牌抽取及交易。

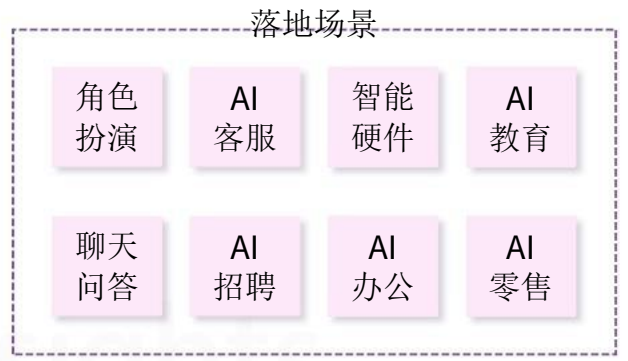
根据 Sensor Tower 数据，Talkie 2024 年 H1 在美国市场累计下载量 380 万次，位列美国同类 AI 产品排名第一，超过 Character.AI。目前 Talkie 全球月活数据已达到 1100 万，一半以上用户来自美国，与 Character.ai 月活数据 1700 万的差距逐渐缩小。

生产力应用



海螺 AI：海螺 AI 于 2024 年 5 月上线，支持多种交互方式：可以发送文字消息、上传图片、PDF、PPT、TXT、DOCX 等格式文件，支持语音通话。8 月下旬最新开放音乐创作与视频创作入口，可由此进入海螺音乐及海螺 AI 创意视频平台，调用海螺视频生成大模型能力。

API 开放平台



MiniMax——融资 & 团队：率先晋独角兽，Under 35 创始人带队

估值高、投门槛高、融资金额高：MiniMax 作为国内大模型创业六小强中唯二在 ChatGPT 爆火之前就成立的大模型公司，很早就吸引了头部机构注意，目前已集齐阿里、红杉、高瓴、经纬、IDG、顺为、明势、云启一众头部机构，估值 25 亿美元。

人才技术实力强，产融经验丰富：创始人兼 CEO 闫俊杰具备丰富的技术、产融商业化管理经验，核心技术团队拥有世界顶尖自然语言处理、语音、计算机视觉、计算机图形学等工业和学术经验。

融资历程

| 时间 | 轮次 | 金额 | 估值 | 机构 |
|---------|-----|---------|------------|--------------------------------|
| 2024/03 | A轮 | 6 亿美元 | 25 亿美元 | 阿里巴巴、红杉中国、高瓴投、经纬创投 |
| 2023/06 | A轮 | 2.5 亿美元 | 12 亿美元 | 腾讯投资、绿洲资本、小红书、顺为资本、高瓴投资、IDG 资本 |
| 2022/07 | A轮 | - | 超 2000 万美元 | IDG 资本、明势资本、高瓴创投、米哈游 |
| 2022/01 | 天使轮 | - | - | 云启资本 |

团队成员



闫俊杰
创始人、CEO



- 中科院自动化所博士，在人工智能顶会及核心期刊发表论文 200 余篇，谷歌学术被引用超 3 万次
- 曾任商汤集团副总裁，从零打造通用计算机视觉模型、深度学习工具链和人脸识别相关的技术体系
- 获得多项重要奖项和荣誉，包括超过 10 次重要国际学术竞赛的冠军，广东省技术发明一等奖，两次吴文俊人工智能二等奖等



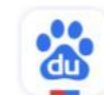
谷焕祎
联合创始人、COO



- 毕业于约翰霍普金斯大学，主修电子计算工程、数学和经济，主研深度学习在医疗影像领域的技术和应用
- 曾任商汤科技融资经理、战略/国际化总监、创新业务总经理，负责集团多项业务孵化、国际化探索，参与集团多轮融资



魏伟
合伙人
开放平台业务负责人



- 曾任腾讯云副总裁，百度技术总监，中国云计算领域的早期贡献者
- 涉足技术和业务两大领域，早在 2011 年就开始使用 GPU 驱动深度学习，支持算力平台的发展

1.1.1 智谱 AI

1.1.4 阶跃星辰

1.1.2 MiniMax

1.1.5 百川智能

1.1.3 月之暗面：
打造爆款 AI 生产力应用

1.1.6 零一万物

月之暗面—模型：明确主打长文本能力，专注迭代文本模型

主打长文本，专注迭代唯一模型：月之暗面立于 2023 年 3 月，10 月首次发布支持 128k 上下文长度的模型时处于业界领先地位，目前 200 万文本长度功能内测中，仅发布唯一一款千亿参数大语言模型，下一阶段即将发力多模态。

AGI 路径战略有一定前瞻性：多次在行业非共识阶段提出公司的 AGI 战略思路，如长文本是必经之路、通过 ToC 产出现数据 Scaling、专注迭代文本能力、在智力密集场景打造超级应用转动数据飞轮反哺模型等观念。



二百万字长文本



根据月之暗面提供的信息，对于中文文本，1 个 Token 大约相当于 1.5-2 个汉字，如果按 1.75 的比例换算，200 万字相当于 114 万 tokens。
(Kimi 智能助手 200 万字上下文目前未对公众开放，只能通过申请内测使用。)



Kimi 智能助手 200 万字上下文功能上传英伟达 2016-2024 年财报后进行长文本分析。

月之暗面—应用：高成本获客打造爆款应用，网感营销风格突出

流量领跑创业玩家，积极探索变现路径：2023年10月推出 kimi 智能助手，2024年3月爆火，带涨 kimi 概念股；6月网页端点击量超越百度文小言；2024年5月，率先上线六档打赏功能，强调灵活性和趣味性，持续积极探索变现路径。

营销能力突出，获客成本高引争议：营销年轻化、网感好，获客手段之一

CPA 高达 30 元，虽然在业内引起争议，但也撮了一个早期爆款应用。

开始通过 API 建设生态，加码 B 端业务：2024年2月正式向企业和开发者开放企业级 API 服务，在增加 B 端服务的同时也开始布局产业生态。

Kimi 智能助手用户流量



根据量子位智库《AI 智能助手 APP 8 月榜单》及《AI 智能助手 8 月 Web 端榜单》，Kimi 智能助手截至 2024 年 8 月 APP 端历史总下载量近 3000 万，位列国内第二，仅次于豆包 AI；Web 端月访问量超 2000 万，位列国内第一。（完整报告请查看量子位智库公众号）

Kimi 智能助手打赏价格

Kimi 智能助手积极探索变现路径，首发打赏制，相比传统订阅制度更加灵活机动，强调趣味互动，营造智能助手伙伴感。提供的高峰期不限速服务平均日价格约 1.2 元/天。海外竞品 OpenAI 订阅价格为 20 美元/月（约人民币 140 元/月），国内竞品大多为免费制。

| 打赏项目 | 价格 | 高峰期使用权限 | 平均每天价格 |
|-------------|-------|---------|--------|
| 送 Kimi 一朵花 | 5.2元 | 4天 | 1.30元 |
| 给 Kimi 加鸡腿 | 9.9元 | 8天 | 1.23元 |
| 请 Kimi 喝咖啡 | 28.8元 | 23天 | 1.25元 |
| 给 Kimi 充充电 | 49.9元 | 40天 | 1.24元 |
| 请 Kimi 吃顿饭 | 99元 | 93天 | 1.06元 |
| 和 Kimi 一起登月 | 399元 | 365天 | 1.09元 |

社交平台网感营销



月之暗面—融资 & 团队：资本宠儿，小天才范本，创业流量担当

融资估值高、金额高、门槛高：截至 2024 年 7 月月之暗面进行 4 轮融资，估值达 30 亿美元，融资金额均为上亿美元，资方全部为头部机构。很大程度由于创始人杨植麟在求学之路早期就开始闪耀天才星，符合资本投“小天才”画像，早早吸引投资人押注，其中部分资方与杨植麟在创办循环智能时就已结缘。创始团队为清华计算机系四同：主要团队清华计算机系本科同。灵魂人物杨植麟标签包括但不限于“天才学霸”、“连续创业”、“顶尖计算机学府”、“懂艺术”、“情智双高”。

月之暗面融资历程

| 时间 | 轮次 | 金额 | 投资方 |
|----------|----|--------|---------------------------------|
| 2024 /05 | B+ | 30亿美元 | 腾讯、高榕创投、鼓手和词曲作者 |
| 2024 /02 | B | 超10亿美元 | 阿里、红杉中国、小红书、美团、砺思资本 |
| 2023 /07 | A | - | 美团龙珠 |
| 2023 /06 | 天使 | 超2亿美元 | 真格基金、毕胜、红杉、清华大学交叉信息研究院助理教授、今日资本 |

创始团队



杨植麟
创始人



- 高中竞赛保送清华大学，大二转去计算机系师从 IEEE Fellow 唐杰，2015 年以年级第一名成绩毕业，本科期间组建摇滚乐队 Splay Tree，各种手机毫秒级人脸解锁技术
- 博士就读于卡内基梅隆大学，在语言技术研究所 LTI 深造，师从苹果 AI 研究负责人 Ruslan Salakhutdinov、谷歌 AI 首席科学家 William Cohen。博士期间，一作身份论文引用次数破万，顶会发表论文 20 余篇，和众多学术明星如 Yoshua Bengio、Yann LeCun、Quoc V. Le 等合著论文
- 博二时和张宇韬创办循环智能



周昕宇
联合创始人



- 清华大学计算机系本科，本科曾与杨植麟共同组建乐队 Splay Tree
- 在旷视科技从事量产工作，与旷视研究院科研负责人张祥雨合作在 CVPR 发表 ShuffleNet 论文，后影响包括苹果 3D 人脸解锁在内的



吴育昕
联合创始人



- 清华大学计算机系计算机视觉方向基础研究，计算机视觉中的检测和识别问题，谷歌学术被引超过 1.9 万次
- 2018 年获欧洲计算机视觉会议 ECCV 最佳论文提名。曾在 GeekPwn 国际安全极客大赛上成为唯一破解人脸识别赛的参赛团队
- 曾在 Meta 人工智能实验室 FAIR，与队员何恺明共同提出组归一化 Group Normalization 方法。期间创建了 Detectron2，Meta 最受欢迎的 AI 项目之一

1.1.1 智谱 AI

1.1.2 MiniMax

1.1.3 月之暗面

1.1.4 阶跃星辰：
惊艳亮相、均衡发力

1.1.5 百川智能

1.1.6 零一万物

阶跃星——模型：模型矩阵能力全面技术路径规划清晰

蛰伏一年，亮相惊艳，Scaling Law 攀登者：2023年4月阶跃星辰成立，一年后首次发布 Step-1 千亿参数语言大模型、Step-1V 千亿参数多模态模型和 Step-2 万亿参数 MoE 语言大模型预览版三款大模型，团队主算力、系统、数据、算力四方全面攀登 Scaling Law 作准备。

MoE、多模态两手抓，迈入万亿模型俱乐部：2024年7月4日，阶跃星在2024年世界人工智能大会上第二次亮相，发布了 Step-2 万亿参数 MoE 语言大模型正式版、Step-1.5V 千亿参数多模态大模型、Step-1X 图像生成大模型三款 Step 系列模型新款。

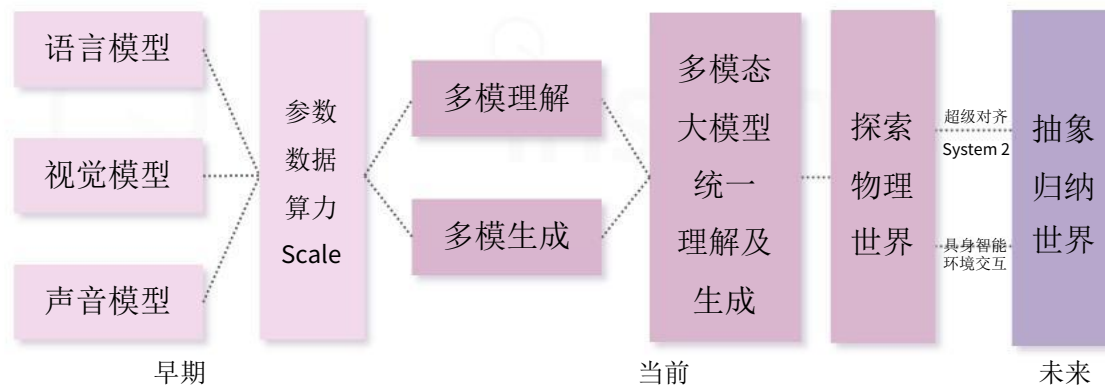
四方面发力攀登万亿参数 MoE 大模型

- 算力** 通过自建机房+租用算力，积极储备算力。
- 系统** 成员实践过单集群万卡以上的系统建设与管理，训练千亿模型的有效算力输出 MFU 达 57%。
- 数据** 数据团队核心骨干来自必应搜索引擎，曾支持全球 100 多种语言，为 200 多个国家和地区提供服务。对全球互联网高质量语料的分布有深入了解。并建立了高质量数据处理和知识图谱流水线。
- 算法** 驾驭多种大模型架构，同时对大模型的认识以及发展路线进行深刻洞察。

模型矩阵

| 类型 | 模型 | 参数 | 架构 | 对标 |
|------|-----------|----|-----|--------|
| 文本 | Step-2 | 万亿 | MoE | GPT-4 |
| 多模态 | Step-1.5V | 千亿 | - | GPT-4V |
| 图像生成 | Step-1X | 8B | DiT | DALL-E |

AGI 路径



主流测试集成绩

| 模型 | 综合 | | | 对战胜率 | | 考试 | | 逻辑推理 | | 数学 | | 知识 | | 编程 |
|-----------------|-----------|------------|---------|----------------------|------------------|--------------|------------------|-------------------|------------------|----------------------|-----------------------|------------------------|--|----|
| | LIVEBENCH | ARENA HARD | MTBENCH | VICUNA (v.s. gpt-4o) | AGEVAL (ch. avg) | BBH (3-shot) | DROP (zero-shot) | GSMBK (zero-shot) | MATH (zero-shot) | MMLU PRO (zero-shot) | MMLU TEST (zero-shot) | HUMAN EVAL (zero-shot) | | |
| Step-2 | 44.1 | 68.1 | 8.9 | 48.3 | 69.5 | 89.4 | 86.0 | 94.0 | 68.4 | 63.0 | 82.9 | 84.1 | | |
| GPT-4-1106 | 45.0 | 74.6 | 9.1 | 41.9 | 63.1 | 80.6 | 79.3 | 90.8 | 64.1 | 68.3 | 85.8 | 86.6 | | |
| Claude 3 Opus | 51.0 | 59.2 | 8.8 | 18.4 | 61.7 | 86.8 | 83.1 | 95.0 | 60.1 | 68.2 | 85.7 | 84.9 | | |
| Llama3 70B Chat | 37.5 | 58.6 | 8.6 | 44.5 | 59.2 | 78.2 | 85.0 | 93.0 | 50.4 | 57.2 | 80.3 | 78.7 | | |

阶跃星 Step-2 在 VICUNA、AGIEVAL、BBH、DROP、MATH 等多个国际主流 Benchmark 上成绩超过 GPT-4-1106、Claude 3 和 Llama 3 70B Chat。

阶跃晨——应用：发力 ToC，繁星划意图布局应用生态

ToC 布局生产力和娱乐陪伴两款产品：2024 年 3 月首发亮相时发布生产力智能助手“跃问”和 AI 内容开放平台应用“冒泡鸭”两款产品跃问目前已实现 AI 联网搜索、多格式文档解析、拍照识图、语音输入等功能。

推出模型开放平台，启动繁星划布局生态：阶跃晨开放平台目前已上线，2024 年 7 月启动繁星划计划，寻找多模态优秀初创企业、团队和独立开发者，提供模型、算力及其他配套支持，补充丰富自身应用生态矩阵。

智能助手应用——跃问

这杯三分糖的奶茶，卡路里是多少？



“这杯三分糖的红茶玛奇朵奶茶的卡路里大约在 349 大卡左右。这个数值可能会因具体的制作过程和配料有所不同,但总体来说,一杯三分糖的红茶玛奇朵奶茶的热量相对较高。如果您正在关注饮食健康或有减肥需求,可能需要适当控制奶茶的摄入量。”

可实现拍照问答、图片理解

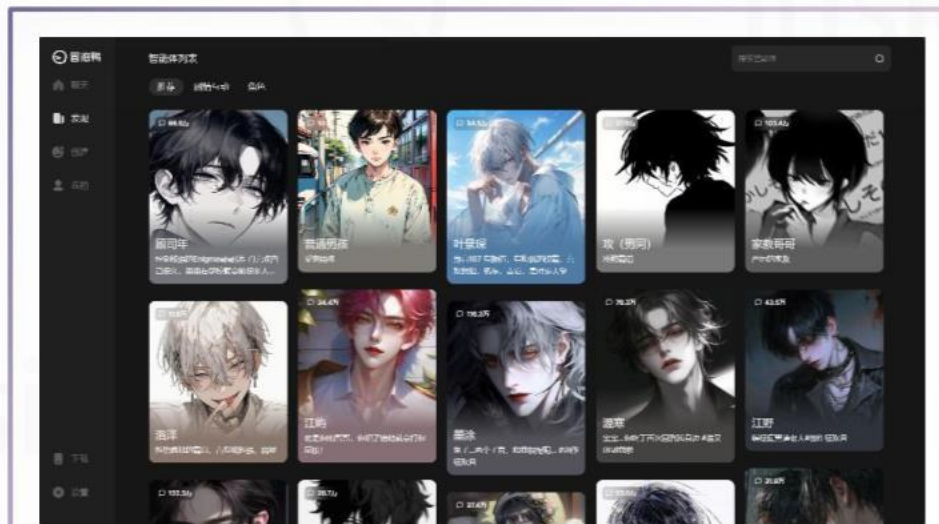
理解方面：目前已支持拍照识图、语音输入、pdf、doc/docx、excel、csv、png、jpg 等多格式文档解析。
生成方面：可以快速完成内容创作、逻辑推理、数据分析等任务。
功能方面：优化了联网搜索、文档解析、推理速度等性能，最新推出浏览器插件。

可实现长篇网页、复杂文档理解及总结

<https://fgk.chinatax.gov.cn/zcfgk/c100012/c5221099/content.html>, 根据这个政策,我最早哪天得完成去年的汇算清缴?

“根据您提供的《国家税务总局政策法规库,即国家税务总局关于办理2023年度个人所得税综合所得汇算清缴的公告,2023年度汇算办理时间为2024年3月1日至6月30日。因此,您最晚需要在2024年6月30日之前完成去年的汇算清缴。”

娱乐陪伴应用——冒泡鸭



冒泡鸭网页端

用户在冒泡鸭上可以根据自身偏好打造智能体和剧情。冒泡鸭支持用户自定义智能体的形象、个性、声音、记忆、背景经历等。

繁星计划

我们支持“星火燎原”

每一个开发者都是一颗闪烁的星火
我们愿提供前行的动力,助你乘风破浪

免费Tokens

赠送总价值4万元的Tokens
助力开发之旅

产品曝光

多渠道推广,让更多人看见你的创意

对接合作

各行业合作伙伴资源对接
实现互利共赢

技术支持

结合项目背景,提供快速
技术支持

免费工位

优先试用地提供
免费工位

上新体验

优先试用最新模型
与功能

RPM/TPM支持

升级为VIP服务,运行更高效

项目落地

优秀的团队和项目,将有机会
获得阶段性的直接投资,
加速落地

目前优秀案例包括胃之书APP、野猫子工作室

阶跃晨——融资 & 团队：团队聚集学术大牛，大模型相关经验丰富

融资进展神秘，资方选择审慎：成立至今未正式披露融资情况，据传仅在 2023 年 11 月完成过一轮融资，投资方或包括红杉中国、启明创投五源资本，据量子位报道，2024 年 6 月阶跃晨正在接洽新一轮估值 20 亿美元的新融资，阿里巴巴在投资机构名单中。

研发实力突出，人才引力强，据传背景深厚：有深厚研发技术积淀，数据算法经验丰富，CEO 技术强，有大团队经验，核心团队聚集来自微软、谷歌、字节、腾讯、旷视等公司相关人才，成立短短一年，规模已达 150 人，先后吸引多位行业大牛加入。据传有著名基金合伙人、独角兽公司创始人、微软华人大佬等支持。



姜大昕
创始人、CEO



- 纽约布法罗州纽约州立大学计算机科学博士，南洋理工大学担任助理教授
- 在机器学习、数据挖掘、自然语言处理、生物信息等领域有丰富经验
- 前微软全球副总裁、微软亚洲互联网工程院 STCA 首席科学家，在微软期间主导多个人工智能项目，曾为 Microsoft 365、Azure、智能语音助手 Cortana 等核心产品提供关键技术支持



朱亦博
系统负责人



- 毕业于美国加州大学圣芭芭拉分校
- 曾任微软研究院研究员，曾任职字节跳动，负责公司 AI 基础设施与基础 AI 框架，积累了万卡集群的搭建经验，曾任谷歌高级主管，负责跟踪、骨架、动作行为等方面研究以及落地工作，曾支持 Anthropic



俞刚
主任研究员



- 山东大学本科，上海交通大学计算机硕士，南洋理工大学博士，曾任腾讯担任研究总监，旷视 Detection 组长，负责检测、分割
- 后在腾讯 PCG 光影研究室从事 CV 相关算法及技术研究



焦斌星
数据负责人



- 中科大和 MSRA 联合培养博士
- 微软必应引擎核心搜索团队前负责人，负责利用数据挖掘和 NLP 算法优化索引和搜索质量
- 曾开发全球高质量站点的自动挖掘算法并用于索引和排序



段楠



- 天津大学博士、中国科学技术大学、西安交通大学兼职博导，天津大学兼职教授
- 曾任微软亚洲研究院 MSRA 资深首席研究员，自然语言计算团队研究经理，主要从事自然语言处理、多模态智能、代码智能、机器推理等研究，多项研究成果成功应用到微软 AI 产品，担任微软 NuWA 系列模型家族项目研究组长和导师
- 谷歌学术被引用次数逼近 1.9 万，与 CEO 姜大昕合著过 44 篇论文



张祥雨



- 西安交通大学与微软亚洲研究院博士生联合培养项目，师从孙剑和何恺明，博士学位
- 曾在微软亚洲研究院工作，期间与何恺明、孙剑、任少卿共同完成了著名的 ResNet，负责底层框架和编码，2016 获得 CVPR 最佳论文奖，2023 未来科学大奖“数学与计算机科学奖”，谷歌学术被引用数已超 23 万
- 曾在旷视工作，与周昕宇共同完成移动端高效卷积神经网络 ShuffleNet

1.1.1 智谱 AI

1.1.4 阶跃星辰

1.1.2 MiniMax

1.1.5 百川智能：
搜狗班底发力 AI 医疗

1.1.3 月之暗面

1.1.6 零一万物

百川智能——模型：开源态突出，模型矩阵全面

开源模型生态建设表现突出：百川智能成立于 2023 年 4 月，6 月发布国内第一款开源模型，8 月首批通过了生成式人工智能备案，榜单中最年轻的公司；

百川大模型支持 LLAMA 热插拔，兼容及可拓展性好，开源下载量领先，累计下载超 700 万次。

搜索引擎技术圈，先天优势：自 2023 年 4 月 10 日成立以来，持续迭代基座大模型，最新一代超千亿参数大模型 Baichuan 4 在 SuperCLUE 评测中处于国内大模型第一梯队，在知识百科、长文本、生成创作等文科类中文任务方面表现优于国外大模型。



开源模型下载量



百川智能 Baichuan2 系列开源模型在开放社区公开累计下载量已经突破 700 万次，最新 hugging face 月下载量超 6 万次。

最新闭源基座大模型主流 Benchmark 成绩

| | 英文通用 | 英文通用 | 中文通用 | 中文通用 | 数学推理 |
|----------------|------|------------|-------|------------|-----------|
| | MMMU | MMBench-EN | CMMMU | MMBench-CN | MathVista |
| GPT4V | 56.8 | 82.1 | 42.5 | 80.1 | 56.5 |
| Baichuan4 | 52.2 | 81.3 | 44.3 | 79.9 | 48 |
| Claude3-sonnet | 47.4 | 67.8 | - | 63.7 | 45 |
| Llava-NeXT-72B | 46.4 | 80.5 | - | - | 46.6 |
| Gemini Pro | 47.9 | 75.2 | - | 74 | 46.5 |

百川智能 Baichuan4 在 MMMU、MMBench-EN、CMMMU 等多个国际主流 Benchmark 上成绩接近 GPT4V。

百川智能——应用：ToC 上线生产力应用，发力 AI 生命医疗

ToC 布局生产力应用，发力 AI 生命医疗方向：百川智能于 2024 年 5 月发布了 ToC 生产力助手百小应，主打多轮搜索能力。另外，还在 2024 WAIC 上发布了 AI 健康顾问应用，具备丰富医学知识，高度拟人，可连续交互并给出诊断建议，百川智能计划在 2025 年实现 L3 级别 AI 医疗。

ToB 以 MaaS 平台为业务重心：百川智能在 ToB 方向推出 MaaS（Model as a service）+ AaaS（Agent as a service）服务，Maas 为重心，分旗舰版和专业版，AaaS 服务已开启公测，API 调用能力接近 OpenAI。

百川智能 ToC 应用



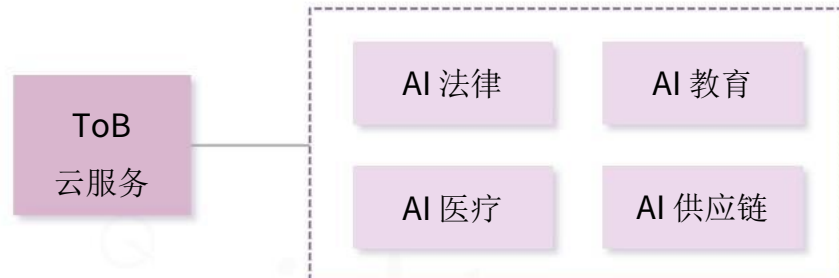
百小应主打多轮搜索功能，相比单轮搜索，在市场调研、产业分析等复杂场景下，能更有效地获取专业深度信息，搜索结果可结构化呈现，还能实现网页链接、PDF、word 文档长文理解分析。

发力 AI 医疗

| 等级 | 百川智能划分 AI 医疗水平 |
|----|---|
| L5 | 完全自动化健康管理 AI 能够管理患者的整个健康旅程，从预防、诊断到治疗，无需人工干预 |
| L4 | 高度自动化诊疗 AI 可以在大多数常见疾病中独立做出诊断和治疗决策，仅复杂案例需医生介入 |
| L3 | 条件自动化诊疗 AI 可以在特定条件下自动推荐治疗方案，但在关键决策时需医生确认 |
| L2 | 多模态辅助 AI 可以整合多种数据源（如病历、影像、实验室结果），提供更全面辅助信息 |
| L1 | 辅助诊断 AI 可以提供数据分析活图像识别等单一功能的辅助，帮助医生做出更好的决策 |
| L0 | 传统医疗 医生负责所有的诊断和治疗决策，没有 AI 介入 |

百川智能计划在 2025 年实现 L3 级别 AI 医疗

ToB 业务



百川智能——融资 & 团队：A轮融资已完成，搜狗核心班底再创业

已完成 A 轮融资，金额大、估值高：截至 2024 年 7 月，百川智能已完成 A 轮融资，7 月最新一轮融资规模 50 亿人民币，资方包括阿里巴巴、腾讯投资、小米集团等机构以及北京市人工智能产业投资、上海人工智能产业投资、深创投国资。目前已经以 200 亿估值启动 B 轮融资。

王小川带队，携搜狗核心班底再创业：作为搜狗前任 CEO，王小川是模型创业六小强中唯一具备创业并上市经历的创始人，他带领搜狗核心班底创立百川智能，团队技术实力不容小觑，由强系统性工程经验人才，语言模型认知优势的研发人员两部分组成。

百川智能融资历程

| 时间 | 轮次 | 金额 | 估值 | 投资机构 |
|---------|-----|---------|--------|--|
| 2024.07 | A | 50 亿人民币 | - | 三七互娱、中金资本、亚投资本、信雅达、卓源洲、好未来、小米集团、黎资本、深创投红点中国、腾讯投资、阿里巴巴、顺为资本、上海人工智能产业股权投资基金合伙企业、临港科创、北京市人工智能产业投资、顺禧、中贝通信集团股份有限公司 |
| 2023.10 | A | 3.5 亿美元 | 19 亿美元 | 阿里巴巴、腾讯投资、小米集团、心资本、愉悦资本、金山软件、顺为资本、好未来、清华控股、深创投、红点中国、慕华科创、淡策资本、卓源亚洲、众为资本 |
| 2023.04 | 天使轮 | 5000 万 | - | 王小川 |

团队成员



王小川
创始人、CEO



搜 狐
SOHU.com

- 清华大学计算机科学与技术专业，拥有工学学士、工学硕士、工程博士及 EMBA 学位
- 与清华大学共建天工智能计算研究院，任联席院长
- 曾任搜狗公司 CEO、董、搜狐高级副总裁兼 CTO，带领搜狗在美国纽交所上市



焦可
联合创始人、产负责人



- 清华大学计算机系本科，中科院计算所硕士



茹立云
联合创始人、运营负责人



曾任智融集团创始人兼 CEO、百度高级产品经理

- 清华大学计算机系博士，浙江省高考状元，清华大学天工智能研究院副院长
- 曾任搜狗公司 COO，负责运营及搜索事业部，人工智能教育公司葡萄智学创始人兼 CEO



洪涛
联合创始人、商业负责人



Tencent

- 清华大学电子工程系本科，长江商学院 EMBA
- 曾任搜狗科技 CMO、腾讯 PCG 商业化 GM

1.1.1 智谱 AI

1.1.4 阶跃星辰

1.1.2 MiniMax

1.1.5 百川智能

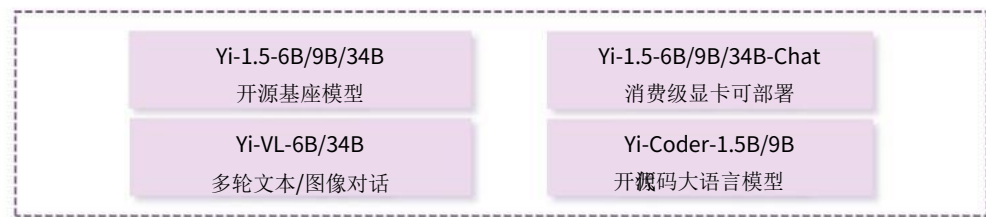
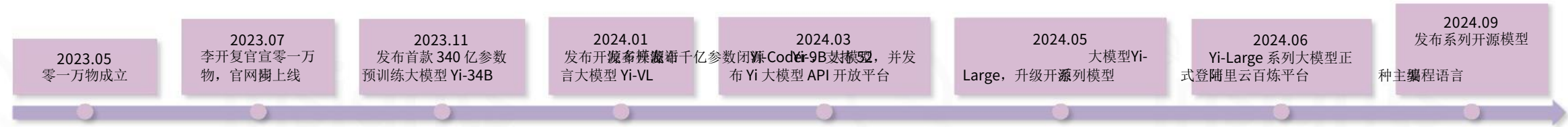
1.1.3 月之暗面

1.1.6 零一万物：
模型首发即封顶开源 No.1

零一万物——模型：AI Infra 技术领先，模型开闭源并重

领先 AI Infra 技术节约训练成本：构建数千卡 GPU 集群进行训练、调优和推理，提高数据有效参数量，以及使用的高质量数据的密度，以构建连续、可控、科学化的训练能力。目前，零一万物具备全栈 AI Infra 能力，在国内率先端到端实现全球先进 FP8 训练框架。

模型取得国际 SOTA，闭源源重：2023 年 11 月首次发布 Yi 系列大模型，涵盖 闭源开源2024 年 5 月闭源模型 Yi-Large 在测试集 LMSYS 中成绩位列国产第一，以千亿参数尺寸追齐硅谷巨头万亿模型。开源矩阵目前已经推出多模态模型、代码模型及通用模型等。零一万物致力于通过自主创新架构加速追赶 OpenAI。



开源模型矩阵

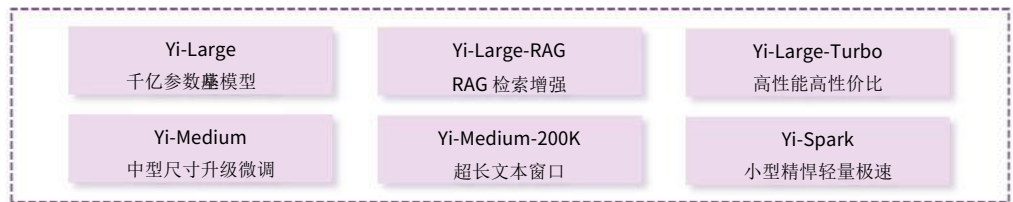
Yi-1.5-9B & 6B Chat Model - as of May 12, 2024

| Model | MMLU | GPQA | MATH | HumanEval | MBPP | MT-Bench | AlpacaEval | AlpacaEval2.0 |
|--------------------------|------|------|------|-----------|------|----------|------------|---------------|
| Yi-1.5-9B-chat | 68.6 | 84.8 | 67.7 | 64.6 | 78.8 | 8.7 | 4.8 | 34.4 |
| Yi-1.5-6B-chat | 62.8 | 78.9 | 40.5 | 64.0 | 70.9 | 7.5 | 6.2 | 17.9 |
| gemma-1.1-7b-it | 60.9 | 57.5 | 24.2 | 41.5 | 55.6 | 7.5 | 5.0 | 7.5 |
| Qwen1.5-7B-Chat | 60.4 | 62.9 | 22.0 | 38.4 | 41.5 | 7.7 | 6.2 | 10.2 |
| Mistral-7B-Instruct-v0.2 | 59.2 | 49.2 | 13.4 | 23.8 | 36.5 | 7.6 | 5.3 | 12.6 |
| Llama-3-8B-Instruct | 65.8 | 54.7 | 21.2 | 61.6 | 61.4 | 8.0 | 6.2 | 20.6 |

Yi-1.5-9B & 6B chat 模型在多个国际主流测试集成绩

Yi-1.5 系列模型大幅提高了数学逻辑及代码能力，在多个国际中英文主流数据集中表现优于同量级模型。

闭源模型矩阵



Yi-Large Global SOTA LLM

Yi-Large Instruction Tuning

| Category | Benchmark | Yi-Large | GPT-4 | Claude3 Sonnet | Gemini 1.5 Pro | Llama3-70B-Instruct |
|-----------------------|--------------------|----------|-------|----------------|----------------|---------------------|
| General | MMLU (5-shot) | 83.8 | 86.4 | 79.0 | 81.9 | 82.0 |
| | GPQA (0-shot) | 43.5 | 35.7 | 40.4 | 41.5 | 39.5 |
| Reasoning | HumanEval (0-shot) | 82.3 | 67.0 | 73.0 | 71.9 | 81.7 |
| | GSM-8K (8-shot) | 92.6 | 92.0 | 92.3 | 91.7 | 93.0 |
| | MATH (4-shot) | 62.4 | 52.9 | 43.1 | 58.5 | 50.4 |
| | MT-Bench | 9.26 | 8.96 | 8.47 | 8.93 | 8.95 |
| Instruction Following | AlignBench | 8.33 | 7.53 | 6.70 | 7.33 | 7.42 |
| | AlpacaEval | 51.9 | 38.1 | 34.9 | - | 34.4 |

Yi-Large 多项国际 SOTA 成绩

零一万物最新闭源模型 Yi-Large 在 HumanEval、MATH、MT-Bench 等数据集上，成绩超越 GPT-4。

零一万物——应用：BG 端多种解决方案，探索 Make AI Work

电商数字人解决方案部署快、成本低：解决方案围绕端侧，主打极简端侧和短视频解决方案，通过构建专属电商语言和声音大模型，实现高质量话术和互动内容。

全行业 ToB 解决方案“一横一纵”打法：ToB 模型解决方案发挥端侧模型横向泛化能力，辅以 SFT 监督学习能力纵向做深做透各行业，提供标准化、垂直精细化全行业 ToB 整体解决方案，目前落地零售、医疗、教育等领域，涉及 AI 搜索、AI 生产力工具、AI 智能巡检等场景。

AI Infra 赋能 BG 端客户：面向 BG 端客户协助政企搭建大模型算力与服务平台设施，加速大模型低成本端到端落地，赋能企业群体、政务服务。



零一万物——融资 & 团队：半年就成独角兽，李开复亲自带队

创新工场孵化，阿里云等多家头部机构押注：作为创新工场塔尖孵化的第七家公司，零一万物可谓“含着汤匙出生”，估值半年就成独角兽。2024年8月，据媒体报道，零一万物已获新一轮数亿美元融资，融资参与方包括某国际战投东南亚财团等多家机构。

七大技术模块招揽顶尖人才领头：零一万物将研发拆解为七大模块技术路线图，各模块均招募顶级科学家和工程师，集合全球最先进的大模型技术认知和论文实践，搭建国际化人才团队。

零一万物融资历程

| 时间 | 轮次 | 金额 | 投资方 |
|---------|----|----------|------------|
| 2024/08 | - | 数亿美元 | 某国际战投东南亚财团 |
| 2023/11 | 天使 | 超 10 亿美元 | 阿里云 |
| 2023/05 | 天使 | - | 创新工场 |

创始团队



李开复
创始人
CEO



机博士

- 曾任腾讯全球副总裁兼大中华区总裁
- 曾任微软公司副总裁，创立微软亚洲研究院
- 曾任苹果公司副总裁



戴宗宏
联合创始人



- 曾任华为云 AI CTO、华为云技术创新部部长
- 阿里巴巴达摩院前 AI Infra 总监
- 曾领导多家主体 AI infra 部门，支持中国三大 100B+ 大型 LLM 模型训练



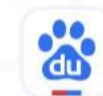
祁瑞峰
联合创始人



- 曾为 Aibee 联合创始人
- 曾任美菜网高级副总裁
- 曾任史宾沙 Spencer Stuart 大中华区科技互联网行业负责人
- 曾任甲骨文、微软和 SAP 前业务高管

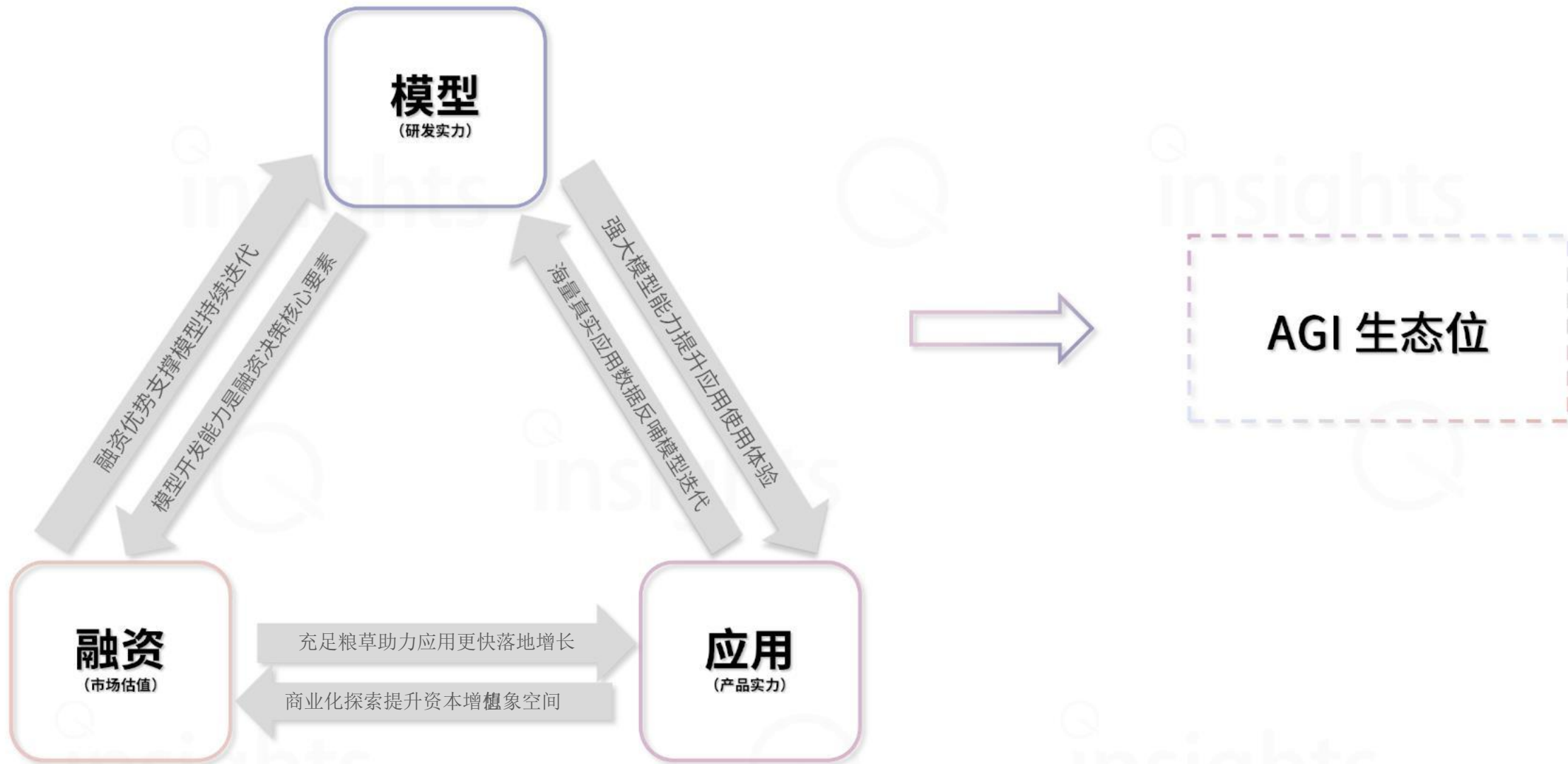


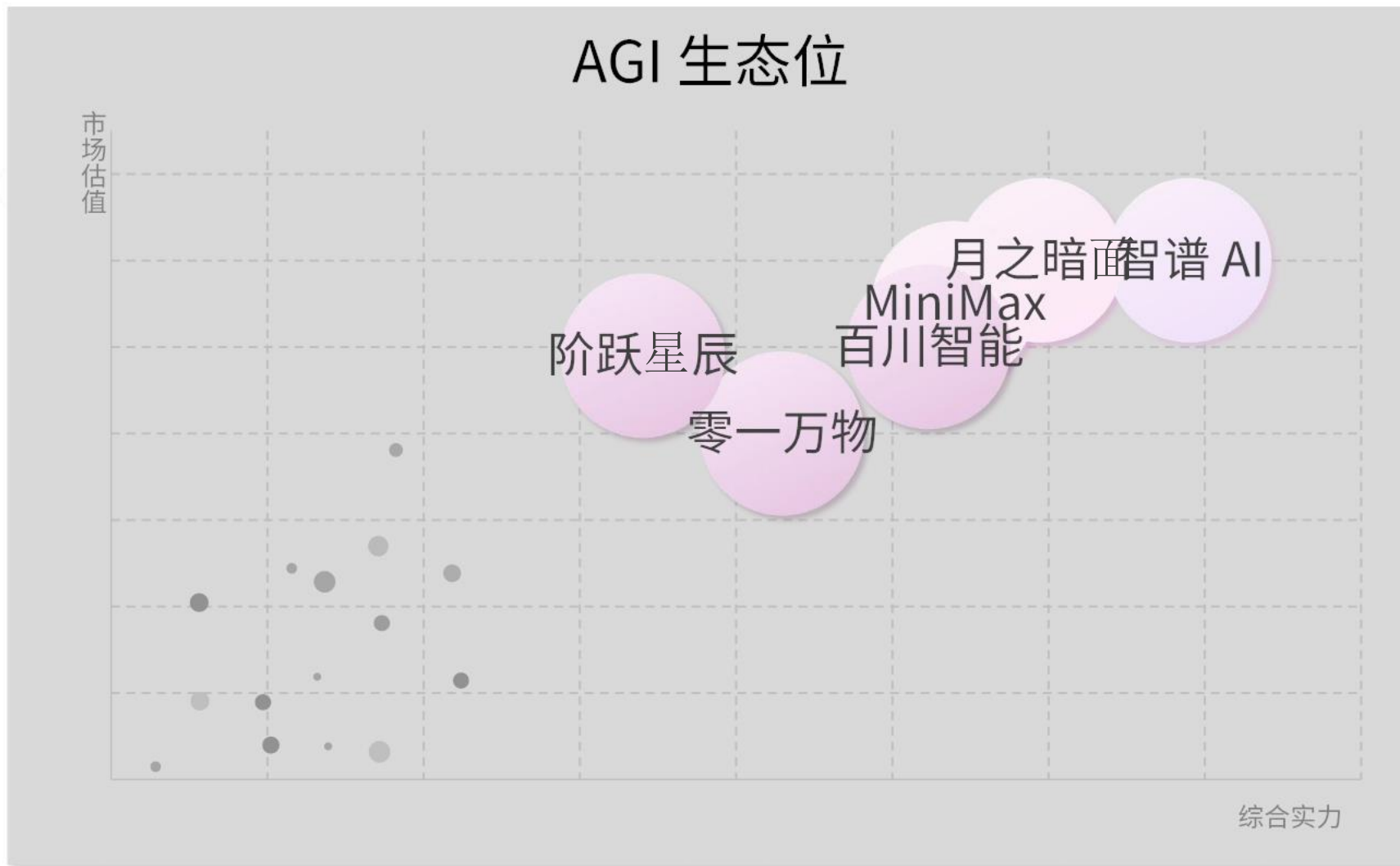
马杰
联合创始人



- 曾任百度安全、IT 和元宇宙副总裁
- 曾为基于 SaaS 的云安全初创公司“安全宝”创始人兼 CEO，后被百度收购
- 曾任瑞星杀毒软件研发主管

创业六小强共贏：转动关键飞轮





大佬派 青年派 学院派

ights

1.2

创业特别关注玩家

insights

1.2.1

DeepSeek:
量化出身万卡持有者

1.2.2

面壁智能

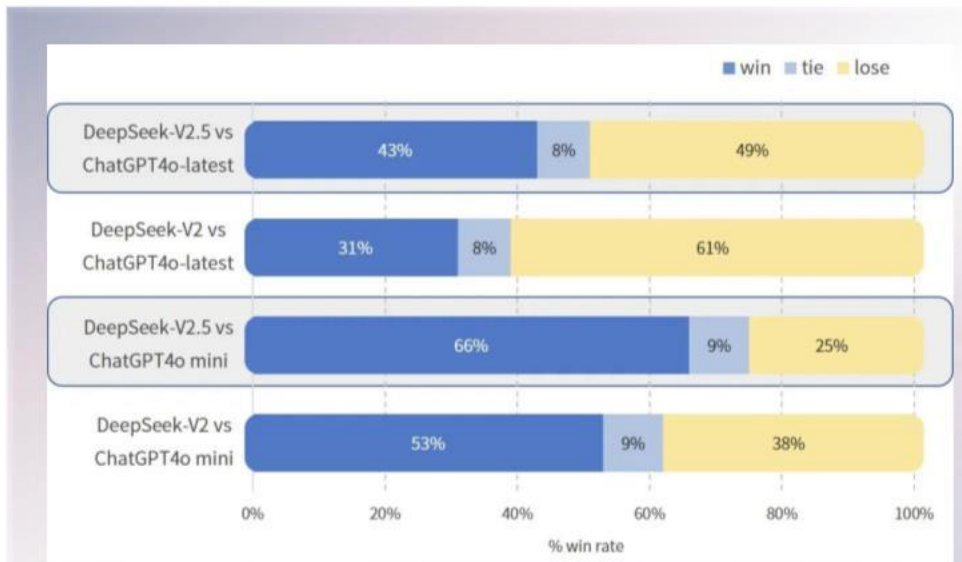
DeepSeek——模型：自研 SOTA 开源 MoE，海外技术传播效果好

脱胎于量化私募机构，自带万卡：DeepSeek 深度求索脱胎于国内知名量化私募机构——幻方量化，幻方量化以人工智能技术进行量化投资而闻名，早年储备万卡算力，不仅如此，幻方量化可为深度求索提供每年几亿的资金支持，在算力和财力上都为深度求索的诞生打下独特基础。

架构创新，模型开源 DeepSeek 坚持进行零到一前沿架构创新和模型开源策略，据创始人表示这源于公司的理想主义圈和技术圈信仰。



最新主力大模型更新情况



LMSYS 大模型竞技场结果

在 DeepSeek 内部中文评测中，DeepSeek-V2.5 和 GPT-4o mini、ChatGPT-4o-latest 的对战胜率（裁判为 GPT-4o）相较于 DeepSeek-V2-0628 均有明显提升。

开源模型矩阵

| 类型 | 模型 | 参数 | 架构 |
|---------|-----------------|---------|---------|
| 通用基座+代码 | DeepSeek-V2.5 | - | MoE |
| MoE | DeepSeek-MoE | 8B/16B | MoE |
| 多模态 | DeepSeek-VL | 7B/1.3B | 混合视觉编码器 |
| 数学推理 | DeepSeek-Math | 7B | MoE |
| 数学定理证明 | DeepSeek-Prover | 7B | MoE |

架构创新

“DeepSeek V2 论文‘可能是今年最好的一篇’。”
——SemiAnalysis 首席分析师

“DeepSeek 论文充满惊人智慧。”
——OpenAI 前员工 Andrew Carr

“这种强烈的反响源自其架构层的创新，国产大模型公司乃至全球开一座大模型都很罕见的尝试。”
——多位行业人士表示

DeepSeek——应用：价格战始作俑，代码生成能力受关注

大厂之外罕见万卡持有者，打响大模型价格战：值得注意的是DeepSeek 的降价不靠烧钱靠模型性能，2024 年 5 月，DeepSeek 发布开源模型 DeepSeek V2 采用先进 MoE 稀疏架构将模型 API 调用成本降至 1 元/百万 tokens，随后字节、腾讯、百度、阿里等大厂纷纷跟进，打响国产大模型价格战。

不过早做应用，代码生成能力引起关注：DeepSeek 目前仅发布网页端 DeepSeek V2.5 通用对话助手、DeepSeek V2.5 代码助手而公众开放模型能力免费使用，不急于发布 ToC 产品追求引领技术前沿，希望与应用创业者共创生态。目前网页端助手的代码能力在行业内引起讨论受到关注。

API 调用价格引发大模型价格战

| Model | API价格/百万Tokens | |
|---------------------|----------------|--------|
| | 输入 (元) | 输出 (元) |
| DeepSeek-V2 | 1 | 2 |
| GPT-4-Turbo-1106 | 72 | 217 |
| GPT-4-0613 | 217 | 434 |
| GPT-3.5 | 11 | 14 |
| Gemini 1.5 Pro | 51 | 152 |
| Claude 3 Opus | 109 | 543 |
| Claude 3 Sonnet | 22 | 109 |
| Claude 3 Haiku | 2 | 9 |
| abab-6.5 (MiniMax) | 30 | 30 |
| abab-6.5s (MiniMax) | 10 | 10 |
| ERNIE-4.0 (文心一言) | 120 | 120 |
| GLM-4 (智谱清言) | 100 | 100 |
| Moonshot-v1 (月之暗面) | 24 | 24 |
| Qwen1.5 72B (通义千问) | 20 | 20 |
| LLaMA 3 70B | 27 | 82 |
| Mixtral 8x22B | 14 | 43 |

DeepSeek 5 月发布的彼时各大模型 API 价格对比

2024 年 5 月，DeepSeek 发布 V2 开源模型，先进架构大幅降低 API 成本至仅 1 元/百万 tokens，是时 GPT-4 Turbo 的七十分之一。随后字节、腾讯、百度、阿里等大厂也纷纷降价，中国大模型价格战由此触发，DeepSeek 也因此被冠以“AI界拼多多”称号。

同名产品



DeepSeek 网页端应用界面

DeepSeek 同名网页端主 作为免费调用模型 力 公 入 存 ， 有 出 小程序，也没有 AI 智能助手类产 能 很多优化。通用对话和代码助手两个功能分别对应 DeepSeek 的通用模型 DeepSeek-V2.5 和代码模型 DeepSeek-Coder-V2.5，通用模型的数学能力和代码模型的编程能力在用户中被讨论和使用较多。

DeepSeek——团队：量化私募出身，团队秘而不宣

脱胎于幻方量化，团队保持神秘：DeepSeek-V2 模型发布后行业内对于其研发团队构成有诸多猜测，从历史来看，DeepSeek 脱胎于幻方量化，或继承其人才圈，而成立后新吸纳人才中据称既没有硅谷海归也没有高管大佬，而本土应届和毕业一两年“没有经验但有潜能”的年轻人才为主，具体人员名单仍保持神秘。

组织架构或利好创新：ChatGPT 没有发生在巨头公司，而诞生于 OpenAI，这让很多人相信 AI 时代的创新需要新的组织架构，DeepSeek 既不侬大厂也不同于纯除初创公司的组织架构也许有利于创新。

创始人



梁文锋
创始人
CEO



- 浙江大学本科、硕士，人工智能专业
- 创立幻方量化，2016 年首次上线运行 AI 策略，到 2017 年底时，投策略全面 AI 化，2019 年资金规模破百亿，2021 年破千亿，后稳居国内量化私募“四巨头”

幻方量化人才画像



幻方量化
圈

- 奥林匹克数学、物理、信息学金银牌得主
- ACM金牌得主
- AI领域大牛和领军人物
- 拓扑学、统计学、运筹学、控制论等学科博士

深度求索人才画像



深度求索
模型团队

- 本土 Top 高校应届毕业生
- 博士实习生

组织架构特点

不拘履历

据创始人梁文锋表示，不同于很多模型公司高薪招聘海外及巨头公司大模型人才，“V2 模型的研发团队没有海归，都本土青年人才，目前 AI 前 50 名顶尖人才也许不在中国，但 DeepSeek 或许可以靠自己打造这样的人才”。

权限自由

据创始人梁文锋表示，“在公司内部，每位员工都可以调动算力和人力，无需审批，同时因为不存在层级和跨部门，员工也可以灵活调用所有人，只要对方也有兴趣。”

松散管理

据创始人梁文锋表示，“DeepSeek 全是自下而上。一般不前置分工，而自然分工。相信每个员工自带想法，无需 push。探索过程中遇到问题，自己会拉人讨论。当一个 idea 显示出潜力时，也会自上而下地调配资源”

1.2.1 DeepSeek

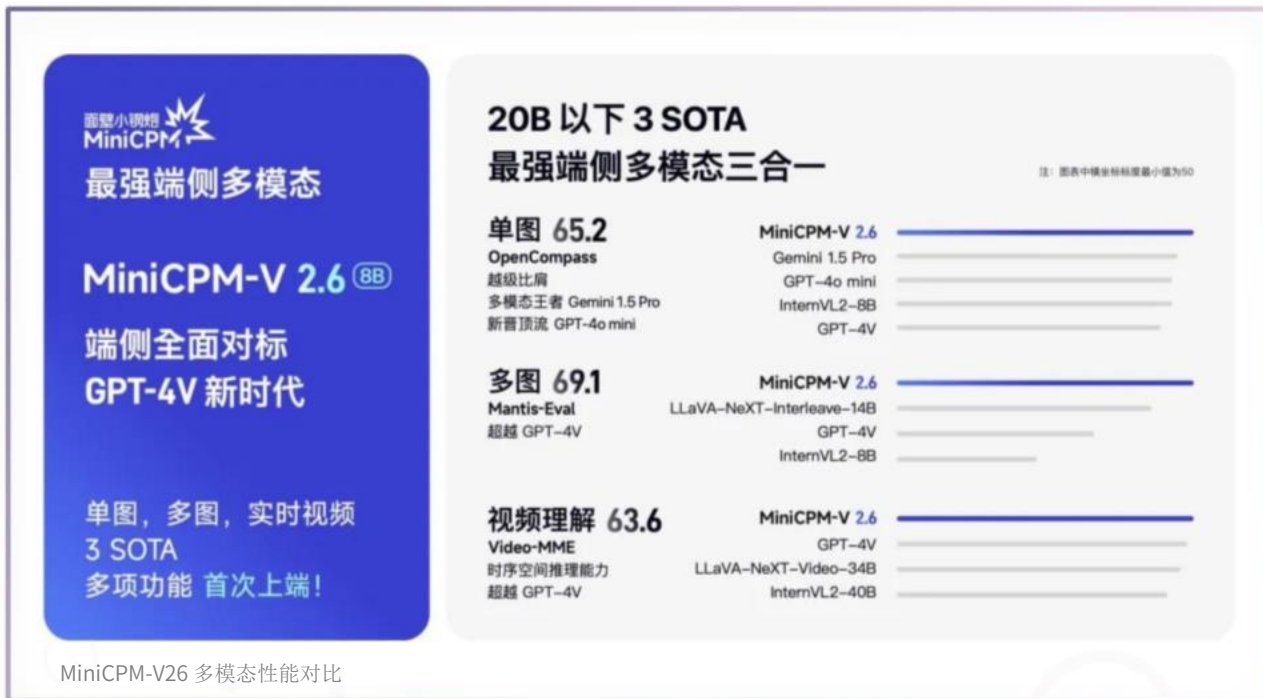
1.2.2 面壁智能： 端侧小钢炮

端侧“小钢炮”性能领先，端云协同探索者：2024年苹果开发者大会后端侧模型概念进入主流视野，国外 OpenAI 和苹果等厂商纷纷发布端侧模型，疆智能在国内很早聚焦端侧，最新开疆模型性能达国际水平，探索端云大小模型协同推理技术，实现模型精细化“降本增效”，在 AI 端侧设备可预见发展后，端侧模型前景乐观。

被斯坦福团队抄袭，意外走红国际：2024年6月，三位美国斯坦福大学学生团队承认开疆模型 Llama3-V 抄袭疆智能开疆模型 MiniCPM-Llama3-V 2.5，两者拥有几乎完全相同的模型架构与代码，疆智能因此意外在国际走红。



最新端侧开源模型性能



面疆智能 2024年8月发布最新小钢炮系列模型 MiniCPM-V 2.6，仅 8B 参数，取得 20B 以下单图、多图、视频理解 3 SOTA 成绩，将端侧 AI 多模态能力拉升至 GPT-4V 水平。

创始人回应抄袭事件



& 团队：脱胎清华，知乎布局，华为投

清华+知乎研发团队：面壁智能名称来源著名科幻小说“三体”，团队脱胎于清华大学 NLP 实验室，聚集杰青欧洲科学院外籍院士、谷歌前员工、智谱 AI 团队前成员等行业优秀人才，CEO 李大海为知乎合伙人、CTO。

已进行天使轮及 A 轮两轮融资：早期天使轮获知乎及智谱 AI 投，A 轮知乎加注，睿创投华为哈勃、北京人工智能产业投资基金加资方。

面壁智能融资历程

| 时间 | 金额 | 估值 | 投资方 |
|---------|-----|-------|----------------------------|
| 2024/04 | A 轮 | - 数亿元 | 春华创投、华为哈勃、北京市人工智能产业投资基金、知乎 |
| 2023/04 | 天使轮 | - 数亿元 | 知乎、智谱华章 |

融资情况



李大海
创始人
CEO



- 北京大学数学系硕士
- 曾为谷歌中国创始员工，云云网工程总监，豌豆荚搜索技术负责人，连续 12 年创业者

曾任知乎联合创始人、CTO，负责整体技术体系搭建搜索推荐业务，知乎 AI 智能社区发起人，帮助知乎实现从百万到亿级的跨越

- 研究方向人工智能和大模型应用



曾国洋
CTO



- 8 岁开始学习编程，高中竞赛金牌保送到清华大学，大二加入清华 NLP 实验室，师从刘知远
- 高三曾在旷视实习，开始接触 AI 领域
- 曾为智源悟道·文源文预训练模型团队骨干成员，研究方向计算机视觉和深度学习
- 作为联合发起人创建 OpenBMB 开源区，模型训练加速和推理加速 BMTrain、BMInf 的主要者之一



刘知远
联创
首席科学家



- 清华大学清华大学计算机系副教授，清华自然语言处理与社会人文计算实验室 THUNLP 成员，清华人工智能研究院模型研究中心副主任，智源科学家

- 期刊和会议发表论文 200 余篇，谷歌学术被引用超 3.1 万次
- 曾获教育部自然科学一等奖，入选国家青年人才计划，《麻省理工科技评论》中国区 35 岁以下科技创新 35 人

ights

2.0

国内巨头玩家格局

insights

国内巨头玩家格局总览



文心大模型



豆包大模型



通义大模型



讯飞星火大模型



天工大模型



混元大模型



快手系列大模型



日日新大模型



盘古大模型

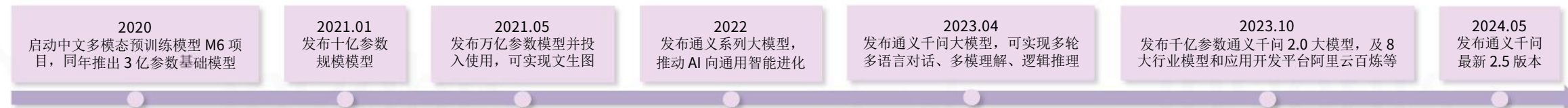
百度：国产 AI 第一重镇，中文大模型头号玩家



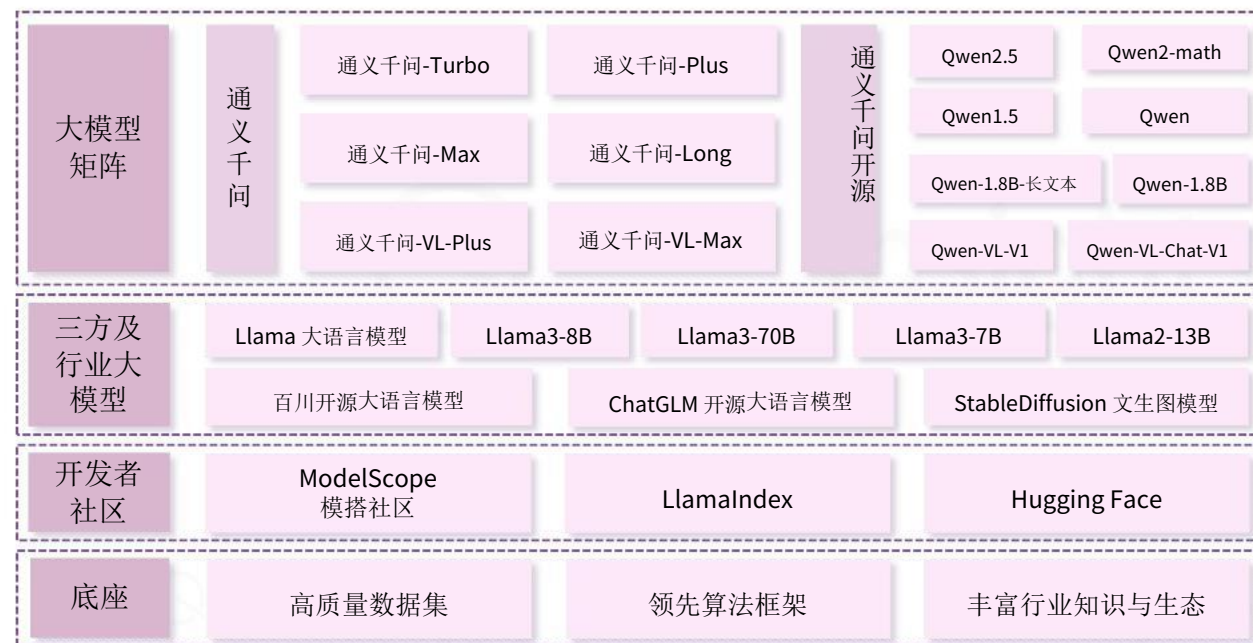
阿里巴巴：投与自研并举，开源模型代表，借助云优势推进 MaaS

洞见 2024 9

MaaS 布局早，服务全 2022 MaaS 并搭建了一套以 AI 模型为核心的云计算和服务架构，积累了丰富的的大模型研发经验、工具和平台，这套能力全部向大模型初创企业和开发者开放，提供包括模型训练、推理、部署、精调、测评、产落地等的全方位服务。



通用大模型



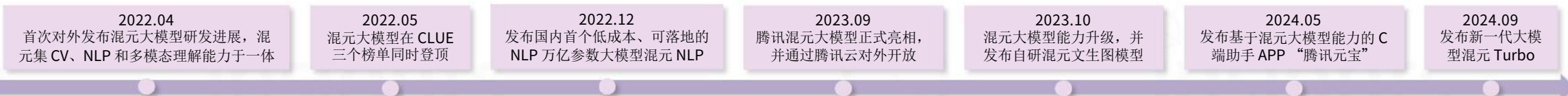
行业大模型

| 名称 | 行业 | 功能定位 |
|------|------|--------------------|
| 通义听悟 | 语音识别 | 生产力助手，提供工作学习语音转写 |
| 通义晓蜜 | 智能客服 | 为企业提供更高效的客服解决方案 |
| 通义智文 | 文本理解 | 帮助用户高效地阅读并理解各类文本 |
| 通义灵码 | 代码编程 | 为开发者提供代码生成、补全等功能 |
| 通义万相 | 图片生成 | 实现图像、视频等多模态理解生成 |
| 通义星辰 | 角色扮演 | 实现娱乐社交场景的个性化虚拟角色创建 |
| 通义法睿 | 法律服务 | 提供法律咨询、合同审查等服务 |
| 通义仁心 | 医疗服务 | 为用户提供健康管理、疾病预防建议 |
| 通义点金 | 金融服务 | 为金融行业提供投资研究、风险分析服务 |

腾讯：全链路自研大模型，ToC 应用靠微信，ToB 落地内外并行

全链路自研，模型训练紧跟前沿：腾讯选择了一条基于自主体系的全链路自研道路，从零开始启动预训练，从 AI 基础设施到机器学习框架，再到模型实现全链路闭环。2023 年 9 月，腾讯发布超千亿参数混元大模型，是业内较早采用 MoE 及 DiT 等先进架构的大模型。

ToC 靠微信，ToB 对内赋能对外开放：ToC 应用元宝基于混元大模型能力，最大特色是深度整合微信生态资源特别是微信公众号这一内容平台。而在 ToB 方面，混元大模型目前已成功应用于广告、搜索和对话等 50 多个腾讯内部业务和产品并通过腾讯云服务对外开放，助力全行业多种应用场景。



元宝+元器

元宝
AI 聊天助手

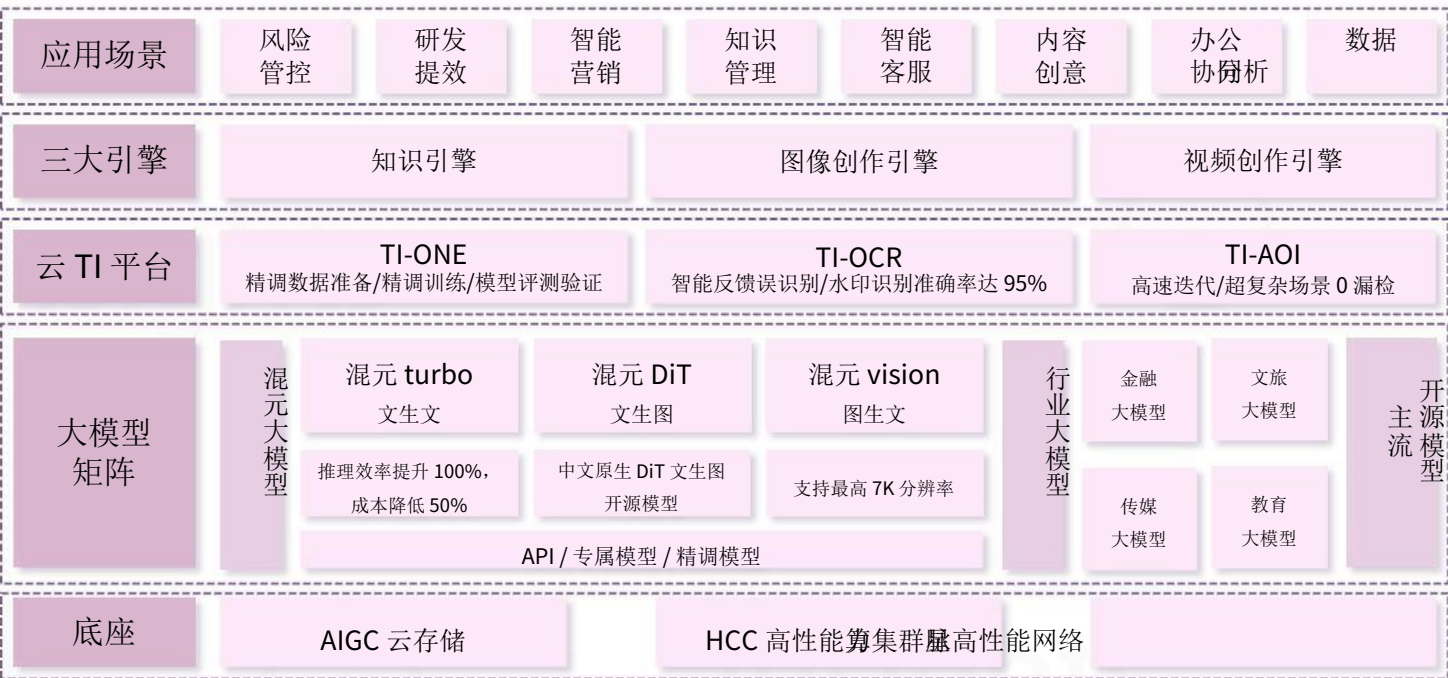
面向普通消费者的 AI 智能助手应用，提供 AI 搜索、总结、写作等日常应用功能。



元器
开发平台

面向企业或开发者，提供一站式 AI 智能体创作与分发平台，支持企业或开发者根据自身需求开发特定的 AI 应用。

模型矩阵及生态

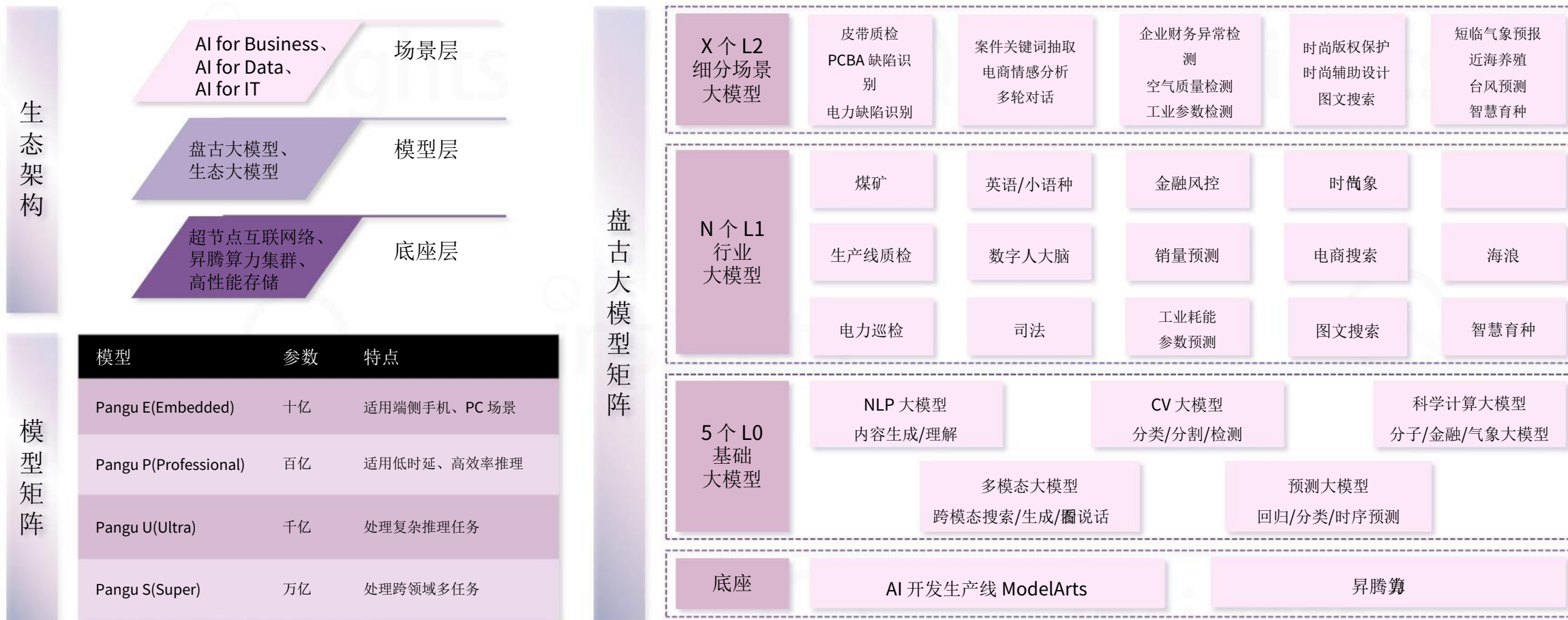


华为：AI算力优势明显，传统行业解决方案广受认可

■的基算底座和 AI 云服务，布局华为云 CloudLake 和 CloudPond 边缘云平台，构建云-网-边-

端协同 AI-native算平台，并提供异构计算架构 CANN、全场景 AI 框架昇思 MindSpore，AI 开发生产线 ModelArts 等，提供英伟达 AI算服务国产替代方案。

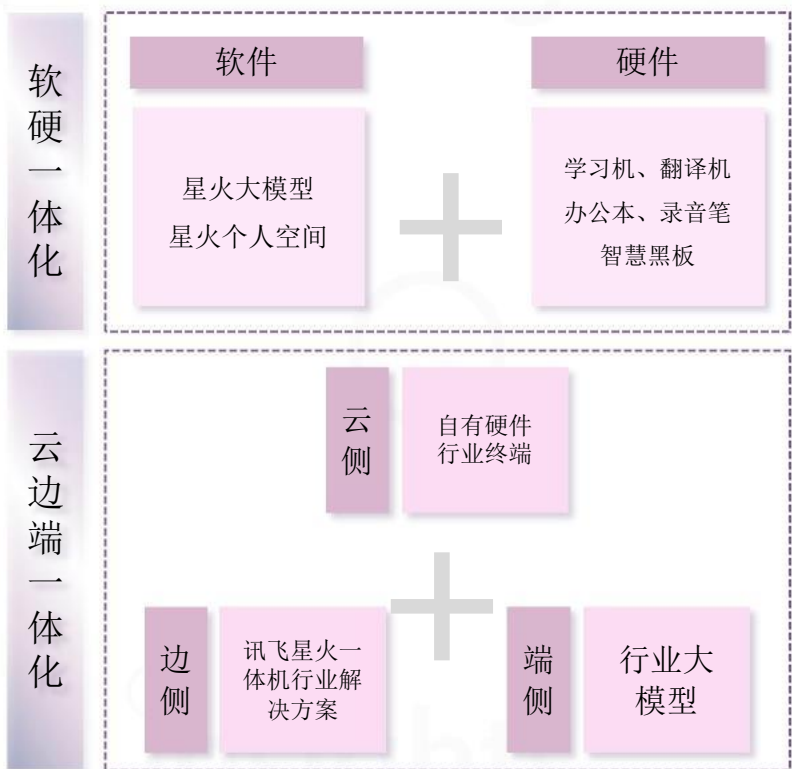
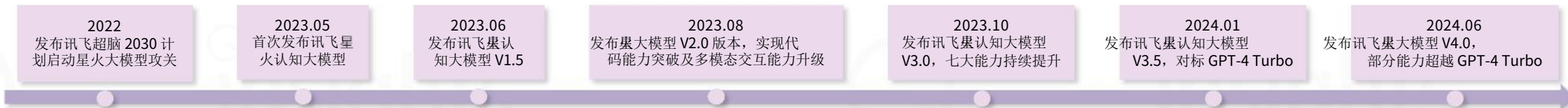
盘古大模型 5.0 专注 ToB，赋能行业：盘古大模型 5.0 分别有 E、P、U、S 四级可选，对应从端侧到跨领域复杂任务的不场景，及文本、视觉、可见光、红外等不同类型数据处理能力，同大幅度提升复杂任务规划能力。目前盘古大模型已经进入到中国 30+ 行业，400+ 应用场景，包括汽车、交通、气候、重工业等。



科大讯飞：国产自主可控代表玩家，云边端软硬件全栈齐备

2024 6月 V4 GPT-4o V4 套

两个一体化布局，G端收紧C端倾斜，财务指标回暖：科大讯飞打出云边端一体化及软硬一体化组合拳，2023年主动收紧ToG业务，向ToC业务倾斜，包括学习机在内的相关业务增长迅速，2024新增相关研发投入超6.5亿元，上半年预计“营收毛利双增”，大模型战略短期暂时未带来明显财务回报，公司盈利能力持续承压。



讯飞星火大模型业务矩阵

| 细分场景 | 功能定位 |
|------|--|
| 企业服务 | 讯飞星火大模型与企业内部知识库相结合，可实现知识问答、智能营销、会议纪要整理以及公文、公函、通知等固定问答格式自动生成等功能，提高企业运营管理效率，解放生产力 |
| 智能硬件 | 讯飞星火大模型打破了传统的基规则的人机交互方式，弥补了机器在知识量和泛化能力方面的不足支持，让手机、电视、VR眼镜等传统的智能硬件，能更好地满足人们日常生活需求。 |
| 智慧政务 | 讯飞星火大模型与政务结合，通过整合政务知识库，协助政府工作人员在公文写作、会议纪要化、知识问答等场景，提升办公效率。 |
| 智慧金融 | 讯飞星火与企业自有知识库结合，可以实现智能问答、智能营销、营销文案生成等功能，帮助企业实现降本增效，增加营销效果 |
| 智慧生活 | 讯飞星火大模型为我们的互联网生活提供了诸多便利，如：实时文本翻译、星座咨询、内容创作、知识问答等 |
| 智慧医疗 | 讯飞星火大模型结合医院自有知识库，通过自助机、虚拟人等方式，为患者提供智能问答和自助流程办理服务，减轻医院导诊人员的负担并提高整体运行效率 |
| 智慧教育 | 讯飞星火大模型在教育领域的应用，如AI学习机、个性化学习手册等，通过 AI 1 对 1 答疑辅导功能、英语作文批改及润色、理科主观题批改等全新功能，提升教育质量和效率。 |
| 智慧汽车 | 讯飞星火大模型在智能汽车领域的应用，如智能车载系统，提供智能导航、语音交互、娱乐服务等功能，提升驾驶体验和安全性。 |

字节跳动：C端豆包数据领先，B端火山引擎改写云计算格局

入场晚，基础好，模型矩阵全面字节到 2023 年 8 月才爆出云雀大模型的消息，到 2024 年 5 月，其他大模型早已多次迭代后，全新的豆包大模型家族才首次亮相，字节一口气发布了包括通用模型、角色扮演模型、语音模型、文生图模型等九大模型，7 月发布图生图模型，9 月底最新发布视频生成模型及同声传译模型。

C 端豆包流量大，B 端火山引擎改写格局：2023 年底字节成立 AI 应用部门，推出多款 AI 应用全覆盖 热门 ToC 应用类型，截至 2024 年 8 月，“豆包 AI” APP 端历史下载已破亿，暂列国产 AI 应用第一。同时，作为市场后来者，B 端的火山引擎平台也积极跟进价格战，以豆包大模型作为核心抓手，打破并改写云计算业格局。

豆包大模型矩阵

| | | |
|-------------|---------------------|------------|
| 通用模型 pro | 通用模型 lite | 角色扮演 模型 |
| 语音识别 模型 | 语音合成 模型 | 声音复刻 模型 |
| 文生图 模型 | Function call 模型 | 向量化 模型 |
| 图生图 模型 | 视频生成 模型 | 同声传译 模型 |

字节是最早跟进大模型价格战的大厂，主力通用模型定价为 0.0008 元/千 Tokens。

2024 年 9 月最新发布的视频生成模型在多动作多主体及多镜头能力上有突出表现，目前已开放企业用户邀测。

豆包大模型应用生态

主力明星应用

APP 历史下载量
1 亿 +

月活跃用户
2600 万 +

APP 智能体
总创建量
800 万 +

豆包
聊天助手



扣子
开发平台



星绘
图片生成



猫箱
娱乐陪伴



河马爱学
教育平台



即梦
视频生成



即创
电商助手

日均 tokens 使用量
1.3 万亿 +

火山引擎改写云计算格局

作为云计算市场后来者，火山引擎抓住大模型机遇成功改写行业格局。依据 IDC 最新发布《中国云服务市场（2023 下半年）跟踪》报告，火山引擎已成中国 GenAI IaaS 领域市场份额最大的服务厂商。

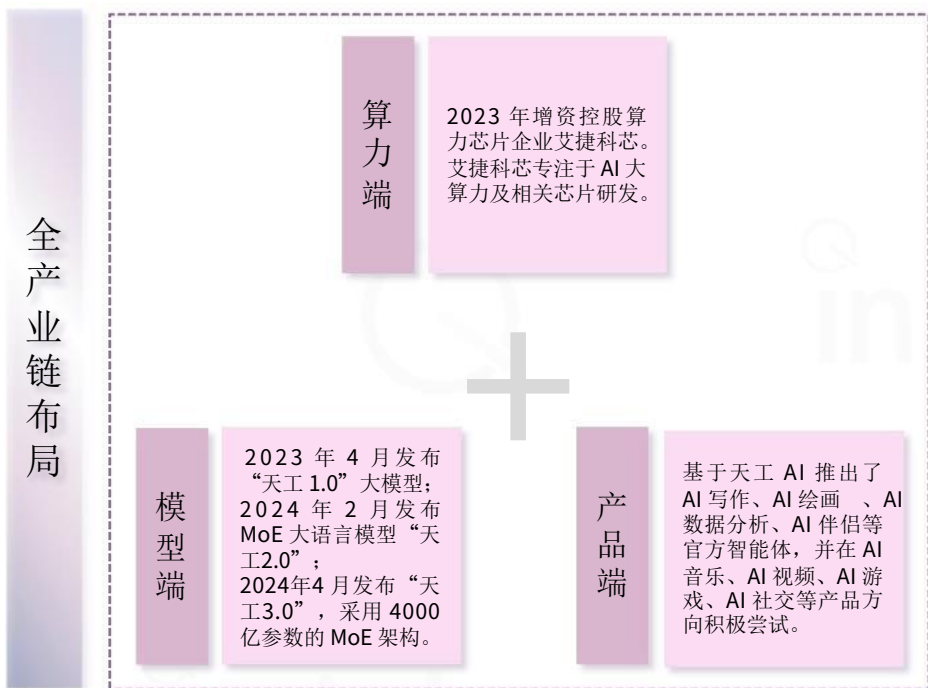
中国 Top5 GenAI IaaS 服务厂商市场份额，2023H2



昆仑万维：浪潮最敏锐跨界玩家，产

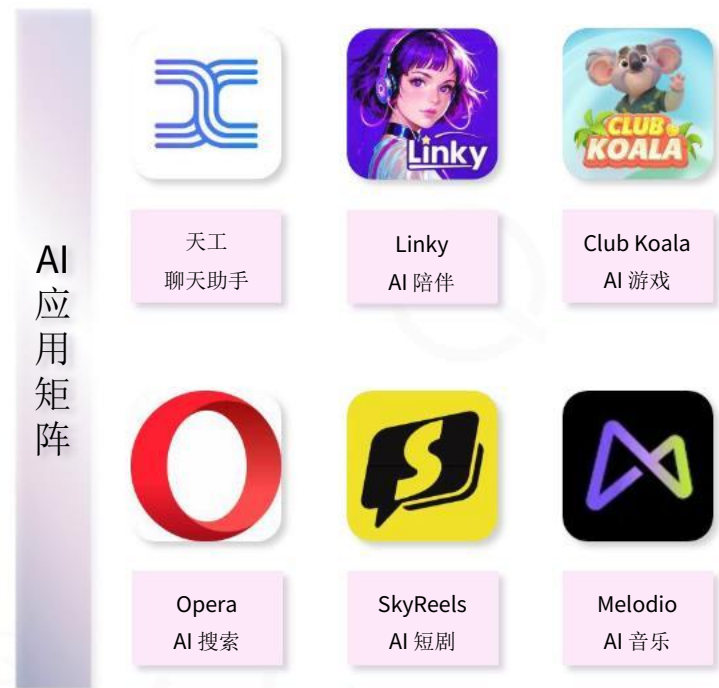
确立“**All in AGI 与 AIGC**”战略，全产业链布局：昆仑万维 2023 年确立“**All in AGI 与 AIGC**”战略，全产业链布局算力端、模型端、应用端，构建了 AI 多元业务矩阵，或成此次 AI 浪潮中嗅觉最敏锐的跨界玩家之一。

ToC 应用形成矩阵、数据表现亮眼：昆仑万维目前已经初步形成包括聊天助手、搜索引擎、短剧平台、社交平台、游戏在内的 AI 应用矩阵。其中，于 2023 年 8 月推出的天工 APP 截至 2024 年 6 月累积下载量已突破千万，DAU 已破百万，APP 端累计下载量次于豆包 AI，在同类应用赛道数据表现突出。



财报业务收入构成

| 业务 | 2023 年财报收入 |
|------|------------|
| 游戏业务 | 5.34 亿元 |
| 广告业务 | 16.5 亿元 |
| 社交业务 | 14.5 亿元 |
| 搜索业务 | 11.4 亿元 |
| 技术收入 | 0.23 亿元 |
| 其他业务 | 1.12 亿元 |



快手：聚焦视觉及推荐，专注业务耦合，AI 视频应用可灵一鸣惊人

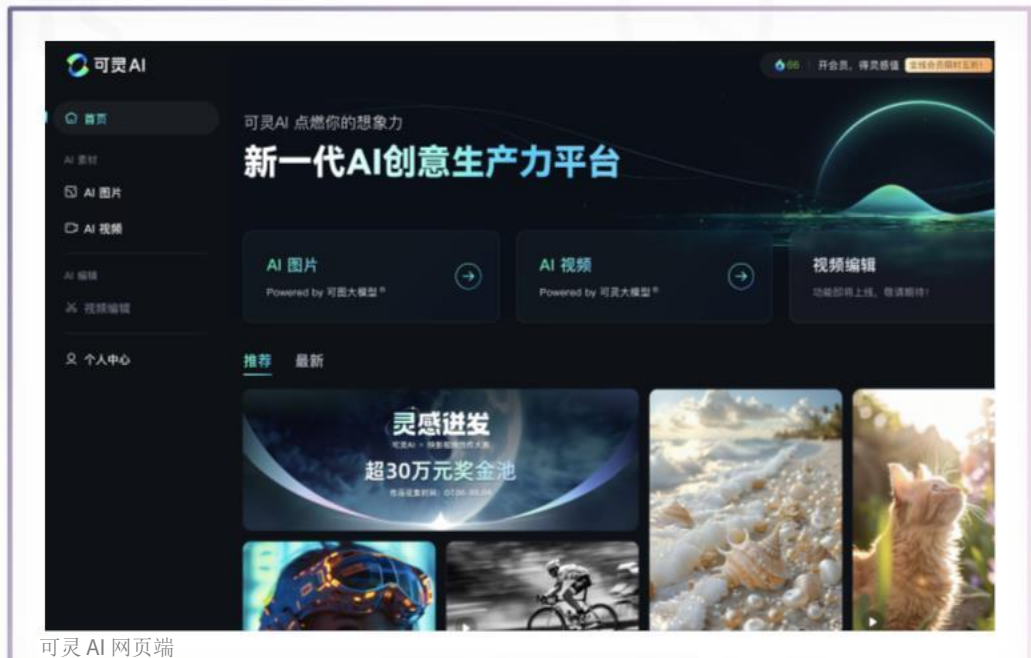
大模型矩阵姗姗来迟，可灵亮相惊艳：2024年7月，快手大模型矩阵及视频生成应用可灵首次亮相。快手搭建了以大语言模型、推荐大模型、视觉生成大模型为核心的矩阵，覆盖内容理解、分发、生成，深度服务快手C端内容业务生态。其中视频生成大模型可灵是国内第一家对标Sora并免费开放使用的产品。

专注ToC业务，赋能内容生产及消费：相比于其他互联网大厂，快手AI战略更加聚焦和落地，不做基建、不卷参数也不对外商业化，而是赋能现有业务，提升生产力和用户体验进行布局。几款大模型和应用主要聚焦两个方向：内容生产和内容消费。2024年7月快手发布国内首部AI短剧——《山海奇境》。

模型矩阵



核心产品——可灵 AI



可灵 AI 网页端

2024年年初Sora引爆互联网后，快手的可灵AI是国内第一家发布产品开放免费试用的视频生成类产品目前视频生成数量超700万，网页端限时免费，可单次生成时长最高可达2分钟。（目前可图大模型能力已合并到可灵AI网页端）

国内首部AI短剧



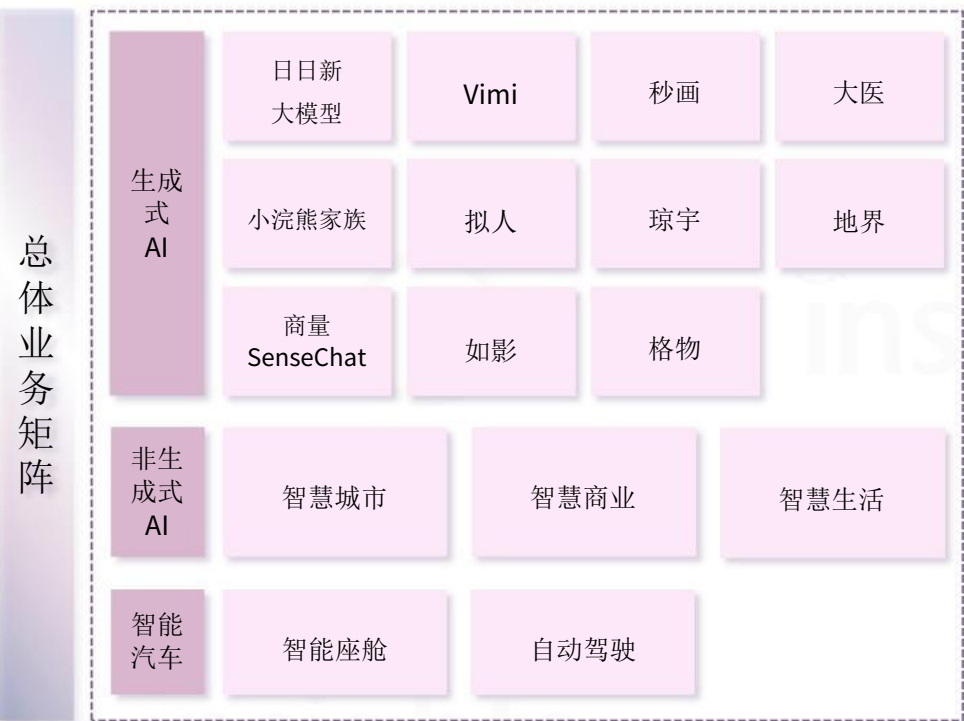
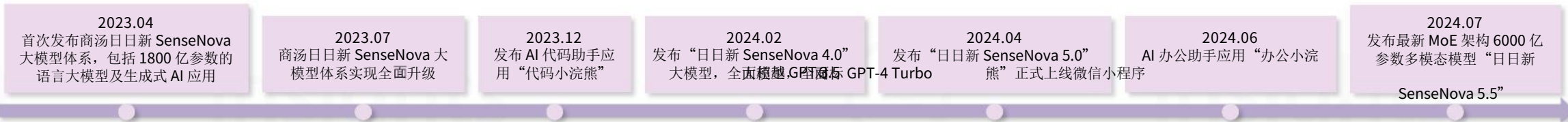
快手 AI 短剧《山海奇境之劈波斩浪》

商汤科技：AI 技术帮助企业实现业务转型，算力模型应用全面局

AI 视觉四小龙实现生成式业务转型：商汤科技起家于 AI1.0 时代的视觉识别技术，生成式 AI 时代来临后积极投身新潮流，发布日日新大模型及系列 BC 端应用，2024 年生成式 AI 业务占比首次超越非生成式业务，成功实现转型。

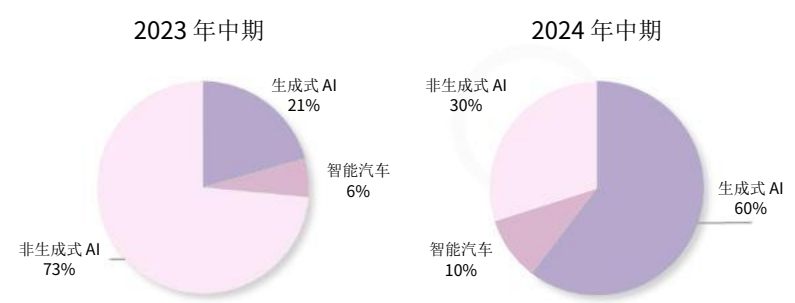
算力模型应用全面局建设，AI 业务矩阵及商汤算力 2018 年就开始自建算力，2024 年 7 月发布最新的 MoE 架构

6000 亿参数多模态模型“日日新 SenseNova 5.5”大模型，对标 GPT-4o，目前业务线主要包括智慧商业、智慧城市、智能汽车、智能生活以及 ToC 应用矩阵。



| 产品 | 场景 |
|-------|-------|
| 代码小浣熊 | 代码助手 |
| 办公小浣熊 | 办公助手 |
| 商量 | 聊天助手 |
| Vimi | 视频生成 |
| 拟人 | 娱乐陪伴 |
| 如影 | 数字人生成 |
| 秒画 | 图像生成 |
| 琼宇 | 三维重建 |
| 格物 | 3D 建模 |
| 大医 | AI 问诊 |
| 地界 | 智能遥感 |

成功转型生成式 AI 业务



2024 年生成式 AI 板块收入近 11 亿元，同比大幅增长 256%，占比首次超越非生成式 AI 业务占比，标志生成式 AI 现已成为商汤科技主导性业务。

ights

3.0

海外玩家格局

insights

海外玩家格局图谱

AI 搜索



具身智能



生物医药



AI Agent & 数字人



代码



其他垂类



通用大模型



视觉生成模型



音频生成模型



硬件



云服务



数据治理



开发平台



ights

4.0

新阶段竞速标准

insights

新阶段 竞速标准

站在 2024 年第三季度的节点上回看，大模型上一阶段的竞争已落下帷幕，而下一阶段的竞速才刚刚开始。

■ scale 方向、高昂的研发投入、行业价格战的挤压和暂未实现的闭环，大模型创业玩家中谁能扛住压力，持续不下牌桌并在新阶段的竞速中突出重围？

有两项不变的重要指标：

第一，**技术研发可持续**，谁能持续推进模型高质量迭代？既靠天才，也需要数据能力。

第二，**商业模式闭环**，谁能率先实现规模化变现？摆脱融资烧钱现状。

上述关键指标，会成为外界对玩家综合实力认定的**准**，有助于玩家占据有利地位，加剧马太效应。此外，大模型浪潮之所以令人激动，核心还是其所代表的底层**范式**机遇，而大模型玩家在完成底层通用大模型研发后，**后续可**能的壁垒和优势，只存在两个维度：

第一，**数据维度**，谁拥有或掌握最高质量和规模的数据？

第二，**应用生态护城河**，对于大模型的竞争力，将开始从性能最强向用户体验最好转移，而用户体验则借助应用、产品和终端触达，卷模型做深，卷开发者生态做广，所谓众行者远，得开发者得决赛入场券。谁能率先建立真正的生态护城河？

6+2 的创业格局，9 个大厂巨头现状，刚刚形成，但也开始面临全新的变数。

量子位 insights

量子位智库

关于量子位智库：

量子位旗下科技创新产业链接平台。致力于提供前沿科技和技术创新领域产学研体系化研究。

量子位，生物计算量子技术及健康医疗等领域最新技术创新进展，提供系统化报告和认知。

通过媒体、社群和线下活动，基于专题技术报道及报告、专项交流会等形式，帮助决策者更早掌握创新风向。

关于量子位：

量子位（QbitAI），专注人工智能领域及前沿科技领域的产业服务平台。

全网订阅超过500万用户，在今日头条、知乎、百家号及各大科技信息平台量子位排名均为科技领域TOP10，内容每天可覆盖数百万人工智能、科技领域从业者。



微信号：Qbitbot020
量子位智库小助手