甲子光年
JAZZYEAR

2024

人工智能开源大模型生态研究

开源为先　场景突破

出品机构：甲子光年智库
研究指导：宋涛
报告撰写：努尔麦麦提·买合木提（小麦）
发布时间：2024.06（初版）
更新时间：2024年6月

目录

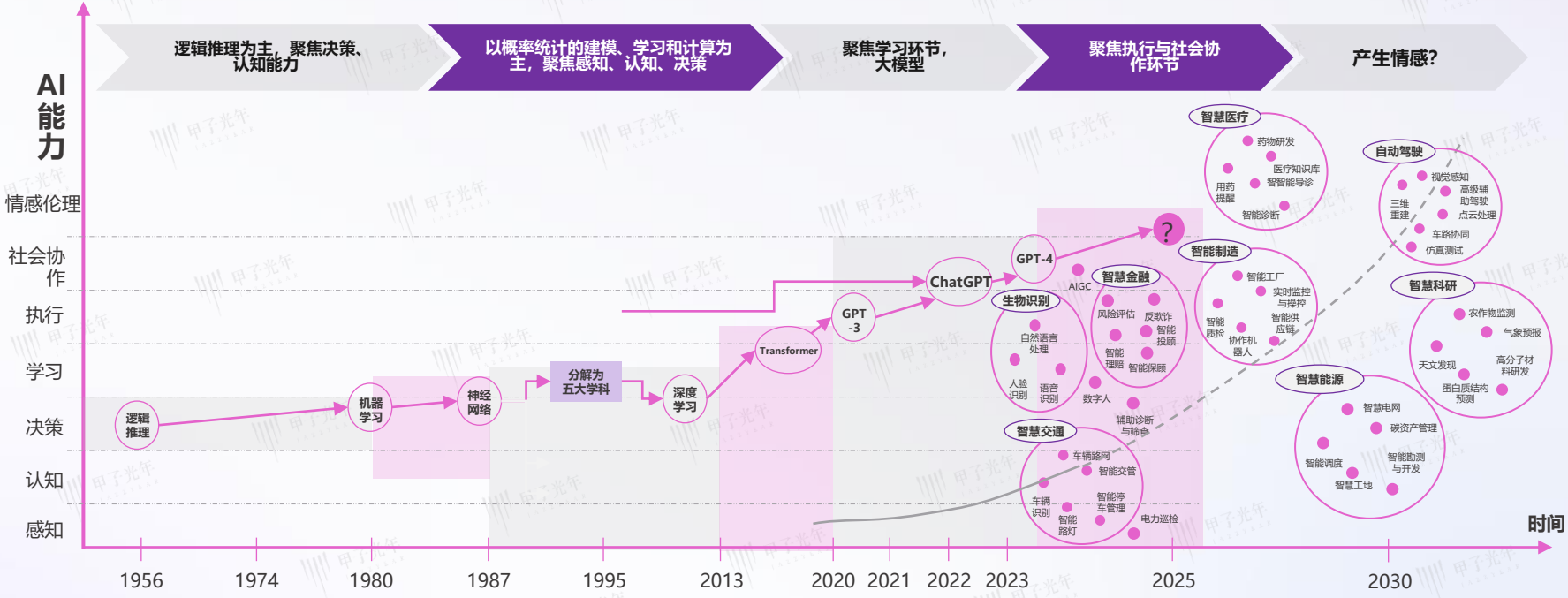
CONTENTS

甲子光年
JAZZYEAR

1.1 人工智能发展进入应用落地阶段

人工智能技术经历70年的发展已经进入成熟期，即将进入大规模应用落地阶段

人工智能即将进入大规模应用落地阶段



数据来源：甲子光年智库，2024年；

1.2 数据、算力、算法作为人工智能发展的核心三要素已经具备基础条件

人工智能三要素：数据（data）、算法（algorithm）和算力（computing power）；

- 人工智能(A)的快速发展依赖于三个核心要素:数据，算法，算力。这个观点已经得到了业界的高度认可。只有这三个要素同时满足了才能加速人工智能的大发展。随着人工智能大模型规模变大以及普及应用，人工智能对能源的需求也在不断加大，逐渐成为人工智能发展关键因素之一。

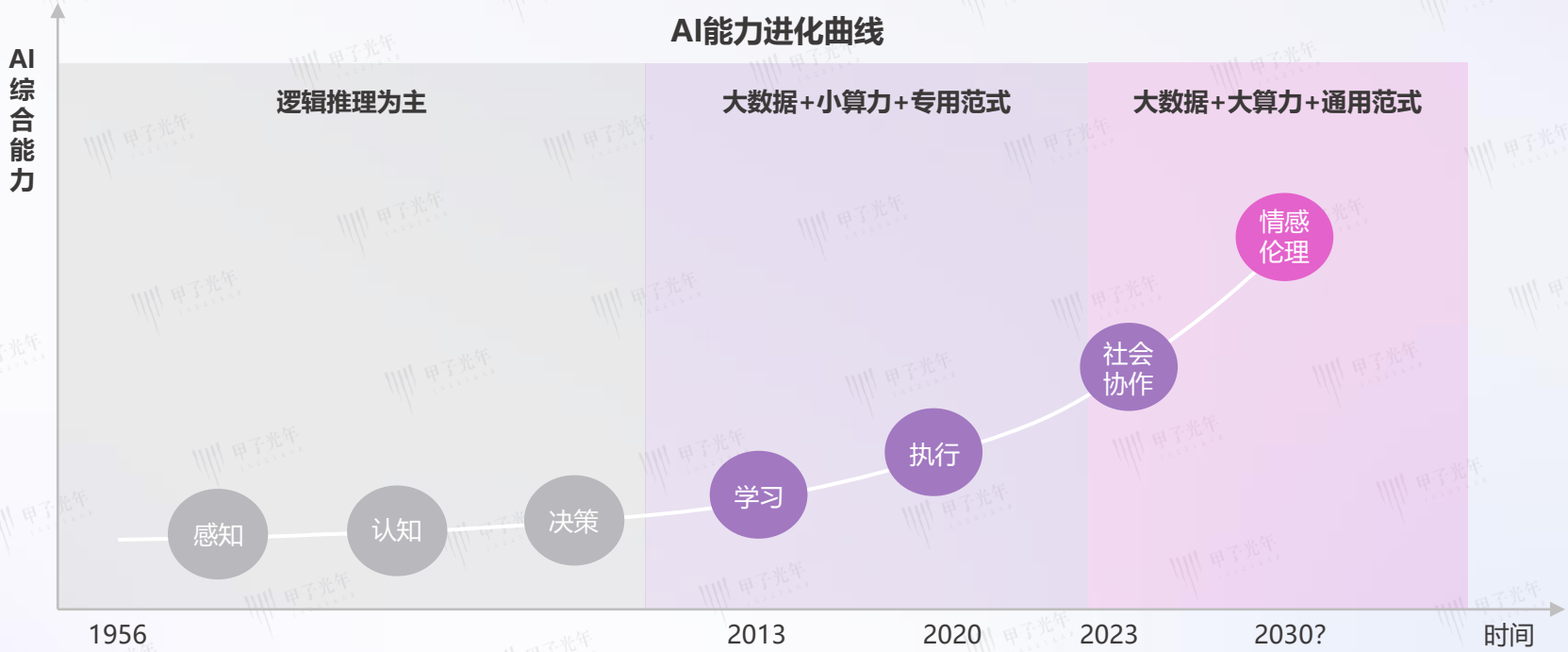
人工智能核心三要素：数据、算力、算法

1.3 大数据+大算力+通用大模型成为新的发展范式

甲子光年 JAZZYEAR

大数据+大算力+通用大模型成为新的发展范式，将推动AI能力提升逼近通用人工智能

持续进化，AI综合能力逼近临界点



数据来源：甲子光年智库，2024年；

1.4 人工智能将推动人类文明生产力跃迁和生产效率的飞跃

工智能智能将推动人类文明生产力的跃迁，标志着人类生产效率出现了第二次脑力效率飞跃

- AI2.0时代将开启社会生产力新变革，**首先体现在对于人类生产效率的颠覆式提升。**
- 人类文明演进依次走过了原始时代、农业时代、工业时代、信息时代、数字时代，到今天的数智时代，每个时代的代表性生产工具都不同。**所有生产工具反映的都是生产效率的提升能力。**数字时代的云、网、端、芯、链等数字工具，除了体力效率的提升之外，还有脑力效率的辅助作用，ChatGPT所代表的AIGC工具的出现，标志着人类生产效率出现了第二次脑力效率飞跃，是新一轮生产力的跃迁，真正实现从体力效率提升向脑力效率提升的转变，这将推动人类社会发生深远变革，其意义不亚于新时代的蒸汽机。

数据来源：甲子光年智库，2024年；

1.5 人工智能进入时代拐点，大模型开源生态成为推动AI产业发展的重要模式

大模型开源生态成为推动人工智能产业从技术走向应用的重要模式

- 开源大模型是指基于开源软件模式，由全球开发者共同参与、共同维护、共同发展的机器学习模型。开源大模型的特点是开放性、共享性和可扩展性，这使得开源大模型在全球范围内得到了广泛的应用和推广。目前，开源大模型已经成为全球人工智能领域的重要发展趋势。
- 模型开源生态不仅加速了人工智能技术的创新，而且推动了其在各个行业的广泛应用。通过开源大模型，企业能够更快地实现任务部署和技术落地，这对于人工智能产业的发展起到了关键作用。随着更多的开源大模型案例和应用的发布，我们可以预见人工智能将在未来的经济社会发展中扮演更加重要的角色。

开源生态的加速形成是大模型时代“安卓时刻”的来临

应用：垂直场景

平台：模型部署

开源大模型

- 开源系统的优势在于影响力的迅速扩散，加快垂直场景应用；
- 开源生态参与者、开发者众多，反应速度快，商业化探索更具潜力；
- 大模型开源，有助于企业/开发者加快实现任务部署和技术落地应用，促进产业发展成熟与生态形成。

数据来源：甲子光年智库，2024年；

目录

CONTENTS

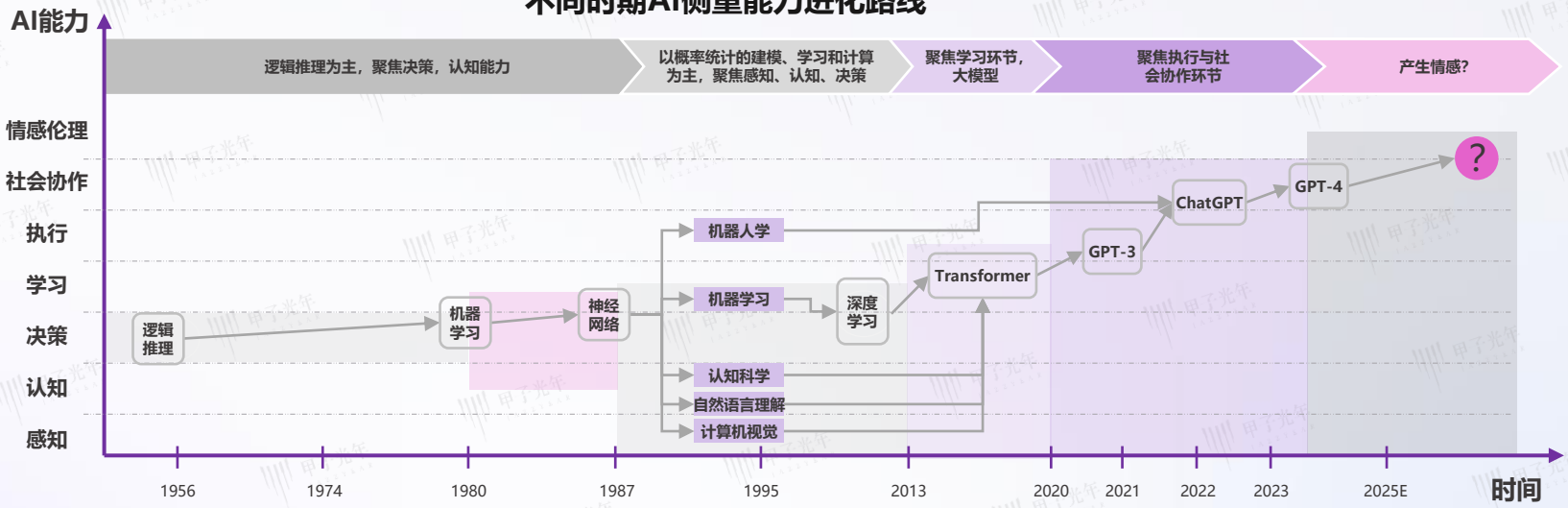
甲子光年
JAZZYEAR

2.1 人工智能技术架构的演变与新趋势

人工智能技术进化出七大核心能力，实现从“解放四肢”到“解放大脑”的升级

- 第一阶段AI以逻辑推理为主，AI能力主要聚焦决策和认知；第二阶段AI注重概率统计的建模、学习和计算，AI能力开始聚焦感知、认知和决策；第三阶段AI聚焦学习环节，注重大模型的建设，AI能力覆盖学习和执行；第四阶段则聚焦执行与社会协作环节，开始注重人机交互协作，注重人类对人工智能的反馈训练。
- 当下正处于第四阶段，这一阶段从2020年开始，代表性事件是GPT-3的发布，突破了以往模型在自然语言处理领域的限制，为语言模型的进一步发展提供了强有力的基础，也为实现智能化的语言交互和人机对话打开了全新的可能性，是人工智能发展的一个关键节点。

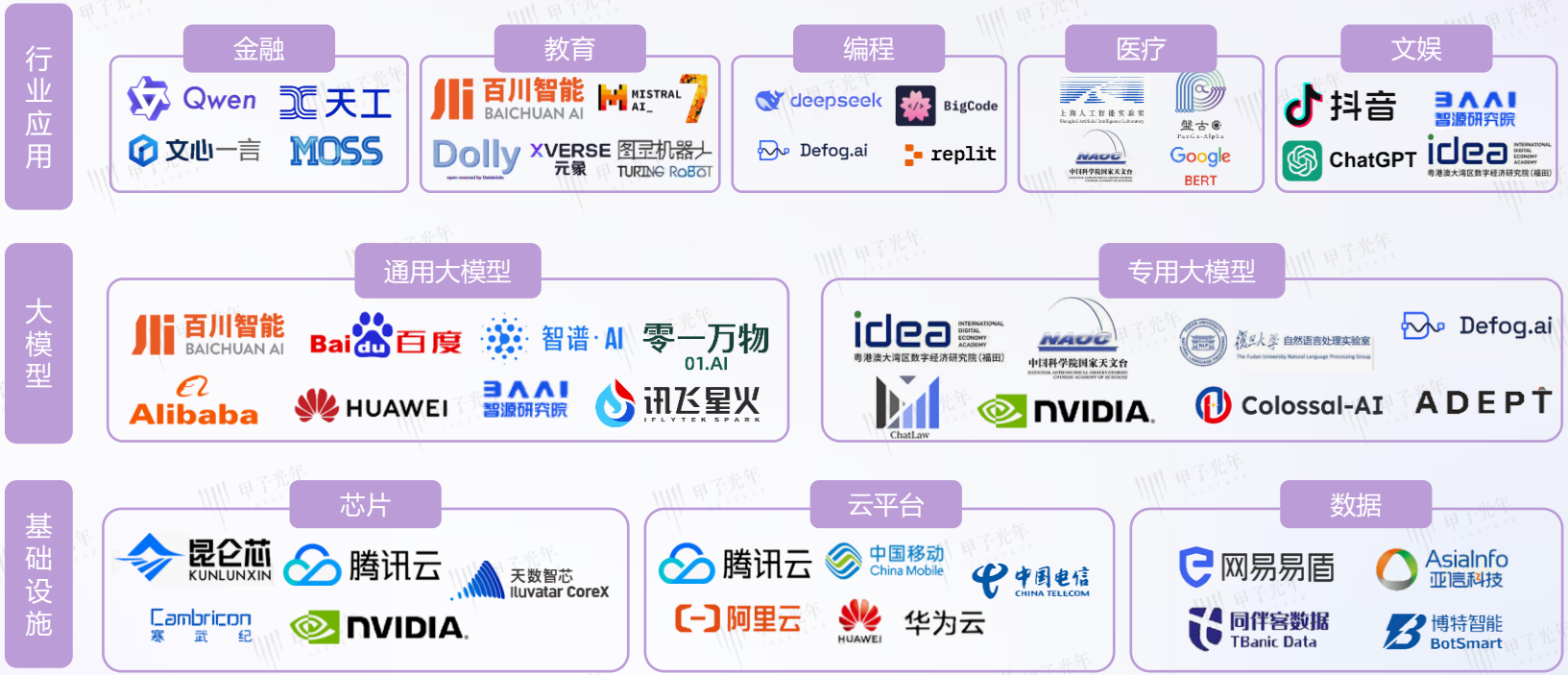
不同时期AI侧重能力进化路线



数据来源：甲子光年智库，2024年；

2.2 基于新一代人工智能开源技术架构的大模型开源生态体系

基础设施、大模型、行业应用构成大模型开源生态体系



数据来源：甲子光年智库，2023；

2.3 大模型开源生态体系的创新主体与创新机制

开源是大模型未来，开源生态体系持续演进

- 开源大模型是基于开源软件模式，由全球开发者共同参与、共同维护、共同发展的机器学习模型。开源由开源规则、开源对象、开源基础设施、参与主体组成。是参与主体在基础设施之上针对对象在遵循一定规则下的一种开放式协作模式，其目的是为了能产生公开复用的产出物。
- 开源的优势，在于降低商业软件采购成本、增强可定制性、保障软件高质量更新、维持技术创新等。

技术流派

- Decoder-only 、encoder-decoder为主流架构
- LLaMa系列单卡版本成为社区热点

所用数据

- 基于Chatbot生成的问答数据集
- 合规高质量数据集

通过Github等平台发布

百川智能 BAICHUAN AI：baichuan-7B大模型已在Hugging Face、Github以及Model Scope平台发布

智谱·AI：ChatGLM开源大模型在Hugging face、GitHub发布

KUNLUN www.kunlun.com：天工Skywork-13B系列模型在GitHub开源

建设自有平台

Alibaba：魔搭社区提供最新最热、开放开源的多领域预训练模型和优质数据集

Baidu 百度：千帆大模型平台不仅提供了包括文心一言底层模型和第三方开源大模型，还提供了各种AI开发工具和整套开发环境

腾讯云：腾讯云TI平台接入LLama2、Falcon等超20个主流模型，支持大模型直接部署调用且可全程低代码操作

数据来源：github，甲子光年智库，2024年；

2.4.1 中国大模型开源生态体系的竞争格局

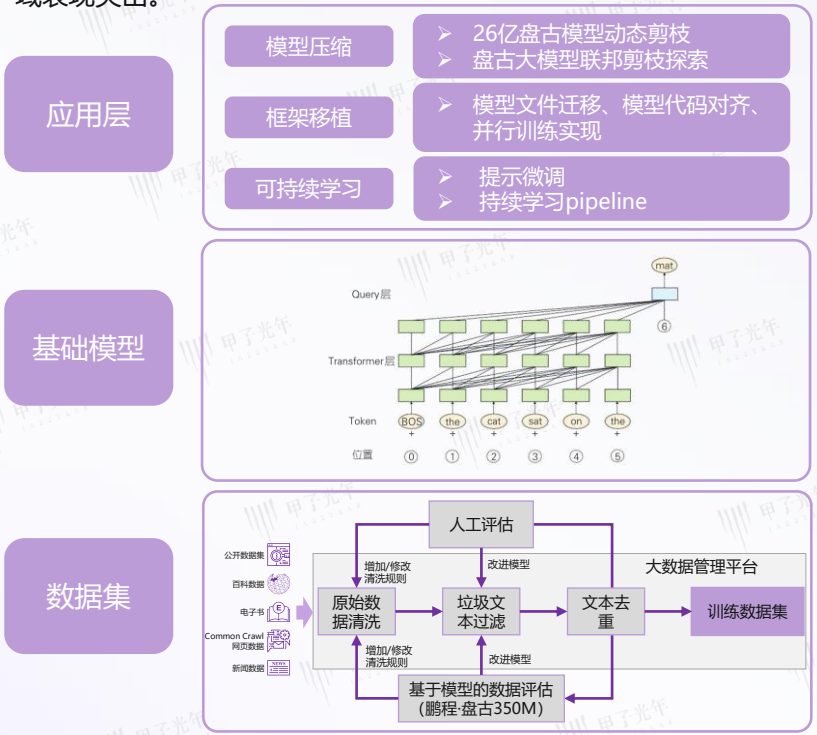
大模型开源生态体系由算力层、基础层、平台层、应用层、安全层构成

数据来源：甲子光年智库，2024年；

2.4.2 中国大模型开源生态体系代表性厂商——华为

鹏程·盘古——大规模自回归中文预训练语言模型

- 鹏程·盘古模型是全球首个全开源2000亿参数的自回归中文预训练语言大模型，在知识问答、知识检索、知识推理、阅读理解等文本生成领域表现突出。



鹏程·盘古模型的规模和参数

模型	参数数量/亿	层数	内层维度	FFN大小	头数
鹏程·盘古 2.6B	26	32	2560	10240	32
鹏程·盘古 13B	131	40	5120	20480	40
鹏程·盘古 200B	2070	64	16384	65536	128

鹏程·盘古模型中文语料数据组成

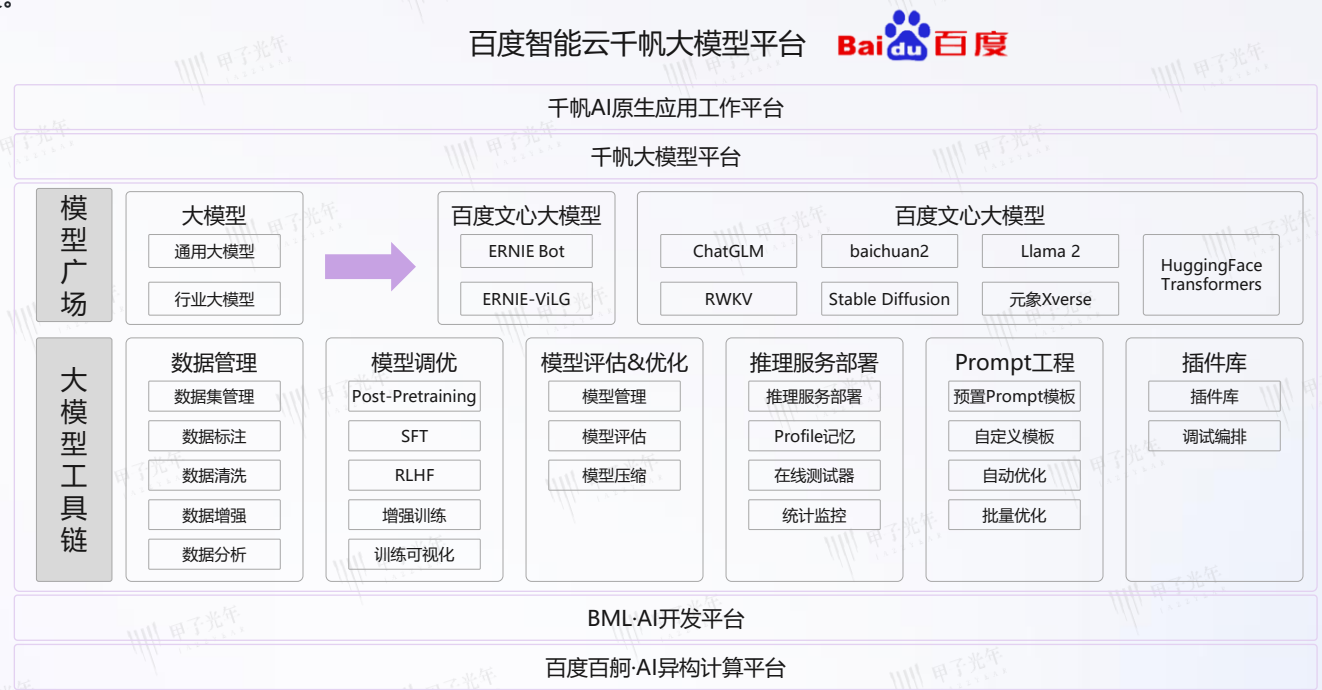
数据来源	大小（GB）	数据源	数据处理步骤
开放数据集	27.9	15 个开放数据集，如 DuReader、BaiDuQA、CAIL2018、Sogou-CA 等	数据格式转换、文本去重
百科数据	22.0	百度百科、搜狗百科等百科类数据	文本去重
电子书籍	299.0	不同主题的电子书籍，如小说、历史、诗歌、古文等	敏感词过滤、基于模型的文本过滤
Common Crawl	714.9	2018 年 1 月—2020 年 12 月的 Common Crawl 网页数据	数据清洗、过滤、去重等所有数据处理步骤
新闻数据	35.5	1992—2011 年的新闻数据	文本去重

数据来源：甲子光年智库，2024年；

2.4.2 中国大模型开源生态体系的代表性厂商——百度

文心大模型——AI应用场景全覆盖

- 文心大模型ERNIE是百度发布的产业级知识增强大模型，涵盖了NLP大模型和跨模态大模型。2019年3月，百度开源了国内首个开源预训练模型文心ERNIE 1.0，此后在语言与跨模态的理解和生成等领域取得一系列技术突破，并对外开源与开放了系列模型，助力大模型研究与产业化应用发展。



数据来源：百度智能云，甲子光年智库，2024年；

2.4.2 中国大模型开源生态体系的代表性厂商——阿里云

通义千问——持续进化的AI大模型

- 通义千问的大语言模型已经实现全尺寸开源——包括18亿、70亿、140亿、720亿7个参数，不同规模和尺寸的模型，可拓宽应用场景。

数据来源：阿里云，甲子光年智库，2024年；

2.5 大模型企业发展面临的问题与困境（1）

大模型训练和应用面临着算力和能耗算力方面的挑战

- 大模型需要大量计算资源，导致全球算力需求指数级增长，对全社会信息基础设施和众多企业、科研机构的大模型研发带来巨大压力。
- 能耗方面，大模型对能源的巨大需求导致人工智能能源消耗占全球能源消耗的3%左右，到2025年将消耗全球15%的电能，给全球环境治理带来挑战。我国大模型发展带来的高能耗可能增加碳达峰、碳中和压力。

01

算力短缺

- 大模型通常需要具有数十亿乃至上万亿个参数，训练时用到数万亿个Token，这就需要消耗巨大的算力。算力需求随着大模型的发展而呈指数级增长，对全球算力规模提出了巨大的要求。大型预训练模型的训练和调优过程需要消耗巨大的算力资源。例如，训练ChatGPT所需的算力相当于64个英伟达A100 GPU训练1年的时间。此外，大模型的日常运营和优化也需要大量的算力投入。预计到2030年，全球算力总规模将达到56ZFlops，其中智能算力成为推动算力增长的主要动力。这对于社会的信息基础设施建设和企业、科研机构的大模型研发都带来了巨大的挑战。
- 根据工信部的数据，2022年全球智能算力中，美国占45%的份额，中国占28%的份额，美国智能算力规模为我国的1.6倍，在中美算力竞争中，我国仍然处于相对劣势的一方。

02

能耗巨大

- 大模型对算力的巨大需求，带来了对能源的巨大消耗。人工智能服务器的功率较普通服务器高6至8倍，训练大模型所需的能耗是常规云工作的3倍。据估计，目前人工智能的能源消耗占全球能源消耗的3%左右，到2025 年，人工智能将消耗全球15%的电能。人工智能的快速发展将对能源消耗和环境产生巨大影响。
- 据估计，GPT-4一次训练的耗电量相当于1200个中国人一年的用电量，仅占模型实际使用时的40%，实际运行阶段将消耗更多能源。一些大型模型运行时的碳排放量巨大，给全球环境治理带来挑战。我国大模型发展的高能耗可能增加碳达峰和碳中和的压力。

数据来源：甲子光年智库，2024年；

2.5 大模型企业发展面临的问题与困境（2）

大模型在数据和资金方面也面临着挑战

- 大模型面临的挑战包括数据获取便利性、数据来源合法性、数据质量可靠性、数据使用安全性、资金投入等方面的挑战。
- 资金投入方面，大模型成本高昂，包括模型开发成本、训练成本、算力成本、数据成本、运维成本等，对普通企业和科研机构而言，资金成为难以逾越的“门槛”。

数据规模与质量待提高

- 数据获取方面，专用类大模型需要专业数据，而这些数据往往属于企业、研究机构等实体，增加了训练难度。
- 数据来源合法性方面，个人信息保护意识的提高使得数据合法使用成为问题。
- 数据质量可靠性方面，开源数据集虽然数量巨大，但质量良莠不齐，从中提取符合预训练要求的高质量数据面临很大挑战。
- 数据使用安全性方面，如何保证使用的数据不带偏见，以及如何保证人工智能制造的数据本身的安全性，都是需要解决的问题。

资金紧缺

- 大模型训练开发成高昂，其成本主要由模型开发成本、训练成本、算力成本、数据成本、运维成本等构成，仅训练成本便动辄高达数百万美元。以Meta大语音模型LLaMA为例，在多达1.4万亿的数据集上，使用2000多个英伟达A100 GPU，训练了21天，花费或高达1000万美元。根据华为公布的消息，开发和训练一次人工智能大模型的成本高达1200万美元。
- 大模型巨大的资金投入，更是将很多小型研究机构和中小型企业拒之门外，导致大模型研发都集中在头部企业和研发机构，加剧了不平等现象。
- 在大模型的投资方面，根据美国斯坦福大学2022年的报告，**美国和中国位列全球投资总额的前两位，但美国的投资是中国的3倍，中国在资金投入方面还有较大差距。**

数据来源：甲子光年智库，2024年；

2.5 大模型企业发展面临的问题与困境（3）

大模型发展在技术和人才方面也面临着挑战

- 针对大模型技术，国内企业与欧美国家存在差距，主要体现在底层架构设计和硬件技术方面。在底层架构设计方面，国内尚无类似的底层架构，大模型的预训练方面只能“在别人的地基上盖房子”；在硬件技术方面，美国占据绝对领先地位，我国自研能力不足，对美国进口依赖程度高，存在“卡脖子”风险。
- 在人才方面，国内大模型人才数量严重不足，与美国相比顶尖人才数量少，制约了大模型研发的快速发展。具体表现为人才数量不足、人才质量不够高和人才外流严重。针对以上挑战，需要加强国内大模型技术的研发，提高自研能力，降低对美国进口的依赖程度；同时，需要加强人才培养，提高人才质量，减少顶尖人才的流失。

技术存在差距

大模型技术涉及软件和硬件两方面：

- 从软件技术看，国内企业与欧美国家存在差距。底层架构设计方面，国内尚无类似谷歌的Transformer模型，对大模型的预训练只能依赖外部技术。在迭代升级和更新换代方面，国内企业也落后于欧美企业，竞争劣势明显。
- 从硬件技术看，在人工智能GPU方面，美国占据绝对领先地位，我国自研能力不足，对进口依赖较高，存在风险。当前大部分大模型训练所用的GPU由美国英伟达公司生产，国产GPU与其性能差距明显。美国已禁止向中国销售A100，而英伟达推出了性能更强的H100，并将优先部署在自家服务器上。

顶尖人才严重不足

国内大模型人才数量严重不足，与美国相比顶尖人才数量少，制约了大模型研发发展。

- 首先，人才数量严重不足。我国人工智能人才缺口超过500万，供需比例严重失衡，人工智能成为“最缺人”的行业。
- 其次，人才质量不够高。与美国相比，国内缺乏顶尖算法人才，数量严重不足。美国在全球最具影响力的人工智能学者榜单中占据主导地位，中国学者数量远远落后。
- 此外，人才外流问题也十分严重。许多国内优秀人才选择出国深造并留在国外，导致顶尖人才的流失。这加大了国内大模型研发与美国的差距，给我国大模型研发带来严峻挑战。

数据来源：甲子光年智库，2024年；

目录

CONTENTS

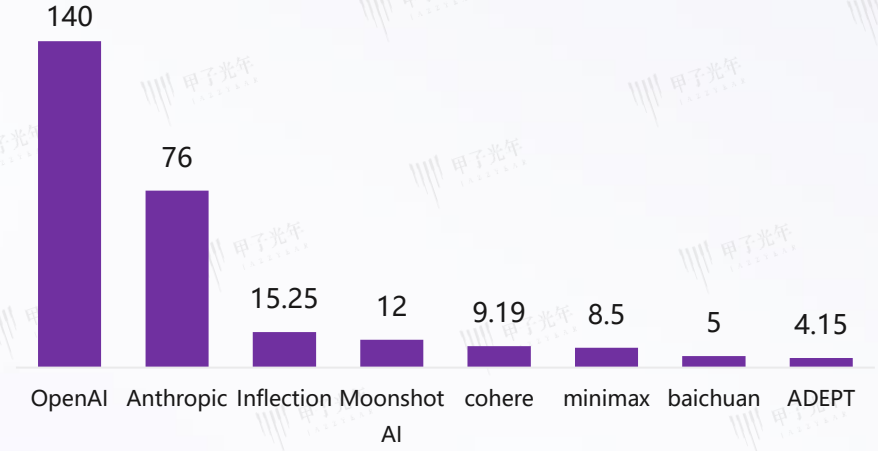
甲子光年
JAZZYEAR

3.1 人工智能开源大模型的投资现状

闭源大模型融资远高于开源大模型融资

大模型资本市场融资情况

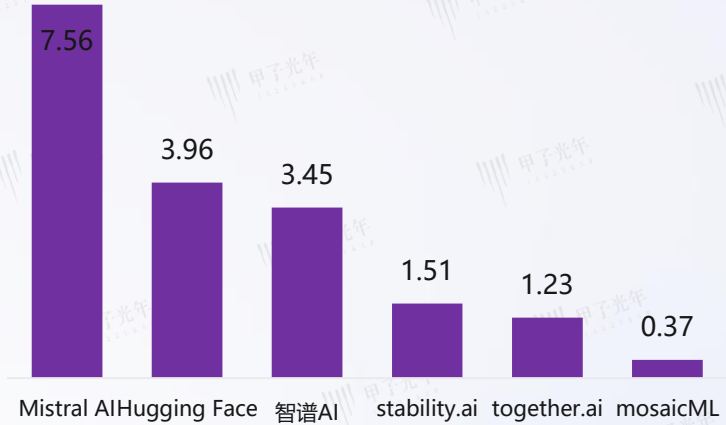
闭源大模型融资规模（亿美元）



*一些开模型厂商可能提供其模型的开源版本，但保留其核心模型的专有权

时间截止：2024年6月13日

开源大模型融资规模（亿美元）



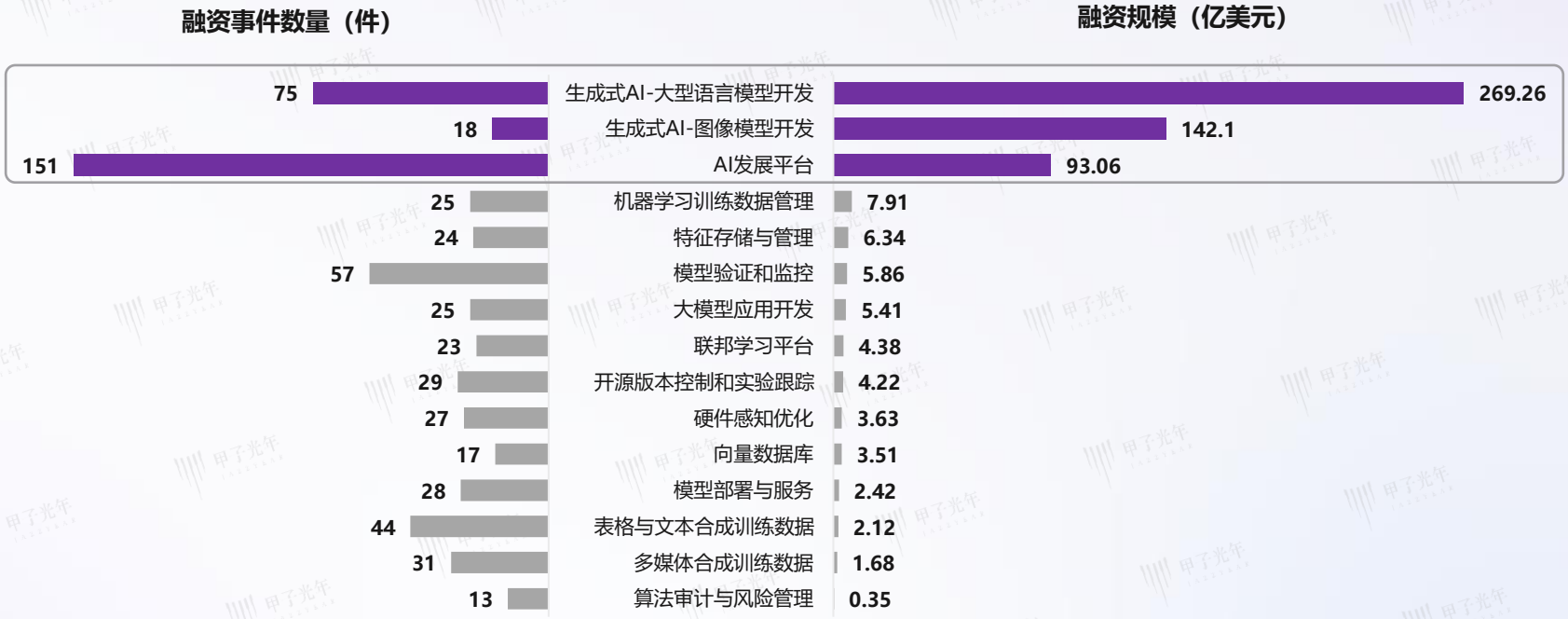
*不包括没有融资的开源开发者

时间截止：2024年6月13日

数据来源：甲子光年智库，2024年；

3.2 人工智能开源大模型的重点投资领域

开源模型总融资事件数量和融资规模



时间截止：2023年10月27日

数据来源：CB Insights，甲子光年智库，2024年；

3.3 开源基金会对推动大模型生态建设的作用

开源基金会将有助于解决模型生态所遇到的挑战，促进AI生态发展完善

- 开源基金会将有助于解决模型生态所遇到的挑战，促进AI生态发展完善。
- 开源基金会可以提供资金、技术、人才等方面的支持，帮助解决模型生态所遇到的挑战，促进AI生态发展完善。促进AI生态发展完善。例如，开源基金会可以资助大模型研发，提供技术支持，吸引顶尖人才，推动大模型技术的发展。
- 开源基金会还可以促进不同企业和研究机构之间的合作，共同解决模型生态所遇到的挑战，推动AI生态的发展和完善。

1 提供技术支持和资源

为大模型开发者和研究者提供技术支持和资源，包括开源工具、框架和库等。这些资源可以帮助开发者更高效地构建和训练大模型，加速生态系统的发展。

2 促进合作与共享

开源基金会鼓励开发者和组织之间的合作与共享。通过共同开发和分享模型、数据集、算法和最佳实践，可以加速大模型的研究和应用，并促进创新。

3 推动标准和规范

推动制定相关的标准和规范，例如模型格式、训练流程和模型评估等方面的标准化。这有助于提高模型的互操作性和可重复性，并促进生态系统的健康发展。

4 保护知识产权和法律支持

提供知识产权保护和法律支持，帮助开发者和组织解决知识产权相关的问题和法律风险，鼓励创新和技术的持续发展。当然，还有其他一些开源基金会在推动大模型生态建设方面发挥的作用

5 资金支持

提供资金支持，通过资助项目、研究和开发者，促进大模型的创新和发展。这些资金可以用于设备采购、研究经费、人员招聘等方面，帮助开发者专注于大模型的研究和应用。

6 安全和隐私保护

鼓励开发者遵循最佳实践，确保模型的安全性和隐私保护。通过提供安全审计、漏洞修复和隐私保护指南等支持，开源基金会帮助保护用户和组织的利益。

7 跨界合作与创新应用

开源基金会鼓励不同领域的交叉合作，例如与学术界、产业界和社会组织等的合作。通过跨界合作，可以加速大模型在各个领域的应用和推广，促进技术的跨界融合和创新。

8 人才教育与培训

提供教育和培训资源，帮助开发者和研究者掌握大模型的相关技术和工具。这有助于提高人才的技术水平和创新能力，推动大模型生态系统的培养和发展。

数据来源：甲子光年智库，2024年；

目录

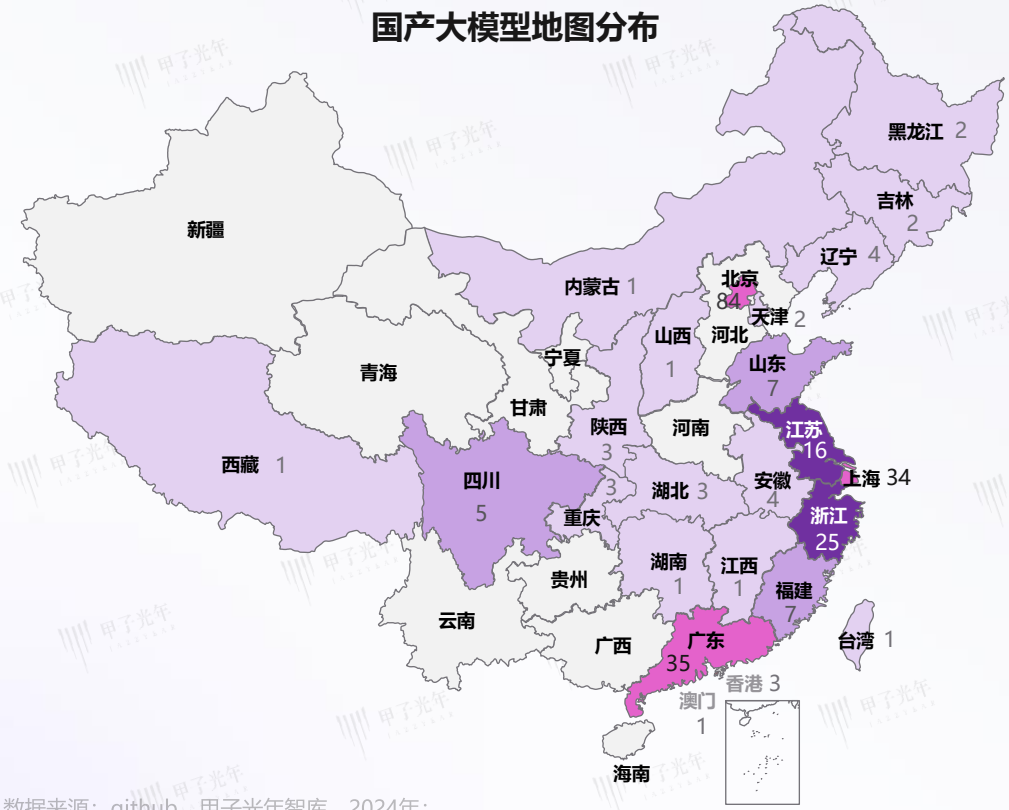
CONTENTS

甲子光年
JAZZYEAR

4.1 大模型产品数量与区域分布情况分析

国产大模型主要分布在北京、长三角和珠三角区域

国产大模型地图分布



国产开源大模型（部分）

北京

- 智普AI：ChatGLM
- 百川智能：baichuan
- 春田知韵（抖音）：BuboGPT
- 面壁智能：CPM-Bee
- 昆仑万维：SkyWork天工，Skywork-MoE
- 浪潮信息：源2.0
- 零一万物：Yi，Yi-1.5 ，Yi-VL
- 智源：智源悟道·天鹰Aquila 7B
- 中科闻歌：雅意2

上海

- 上海AILab：书生·浦语，OpenMEDLab
- 复旦大学：MOSS

浙江

- 阿里巴巴：Qwen，Qwen-1.5，Qwen-1.5-110B
- 深度求索：Deepseek Coder

广东

- 腾讯：Hunyuan-DiT
- 元象：XVERSE

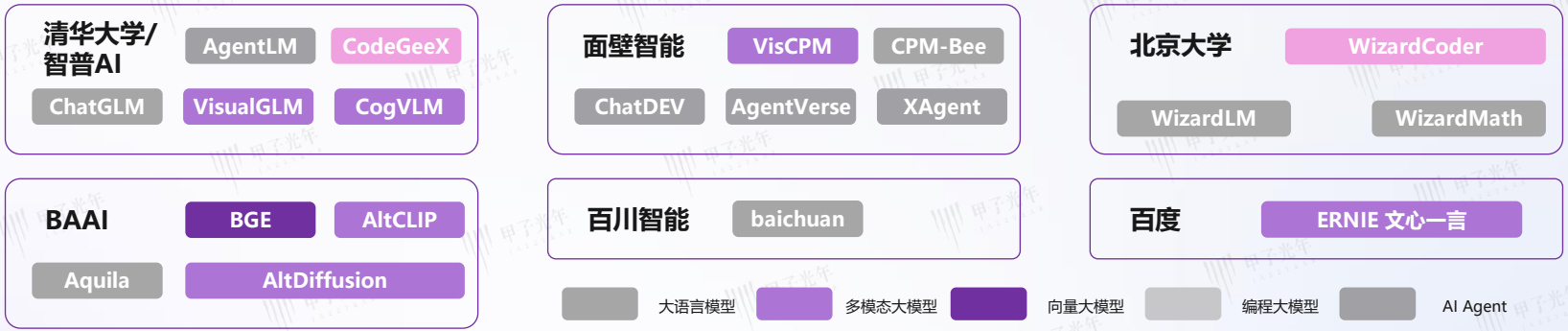
数据来源：github，甲子光年智库，2024年；

4.2.1 北京大模型开源大模型生态发展情况

北京占中国大模型市场的半壁江山

- 据统计，截至2024年6月，我国10 亿参数规模以上的大模型厂商及高校院所共计 254 家，分布于20 余省市/地区，其中北京有 122 家，数量居全国首位，约占全国的一半按模型类型分析，北京拥有通用大模型厂商及高校院所37 家，占比 30%，以百度、智谱华章、百川智能等为代表;行业大模型 85 家，以第四范式、云知声、远科技等为代表。
- 北京大模型的厂商及高校院所可大致分为四类:
  - 人工智能领域的头部企业，以百度、抖音、360 等为代表，在数据、技术、工程化、场景、资金等多方面具备优势。
  - 人工智能领域的高校和科研机构，清华大学、智源研究院、中国科学院等单位的基础研究实力强，聚焦技术创新引领。
  - 人工智能领域的独角兽企业和初创公司，其中智谱华章、云知声、旷视等 AI 独角兽企业，已跑通自研大模型的闭环全流程，可提供 MaaS 模式的 AI解决方案：百川智能、零一万物、衔远科技等 AI 大模型初创公司，迅速布局入场激发大模型创新活力。
  - 传统大数据系统开发企业，以拓尔思、中科闻歌为代表，通过其行业数据积累，推出面向媒体、金融、政务等领域的定制化行业大模型，率先抢占行业应用市场。

北京开源大模型领域典型企业/机构



数据来源：甲子光年智库，2024年；

4.2.2 北京大模型开源社区的典型经验分析

北京是国内开源大模型生态发展较为领先的区域

社区合作与共享经验

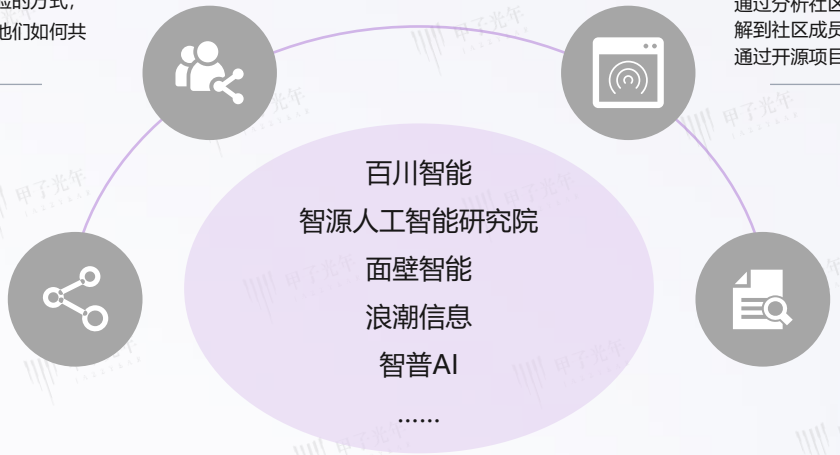
大模型开源社区的核心理念是合作和共享。社区成员可以通过协作开发项目、分享经验和解决问题来共同推动大模型技术的发展。通过分析社区成员之间的合作模式和共享经验的方式，可以了解到社区成员之间的互动和协作方式，以及他们如何共同推动大模型开源社区的发展。

开源项目和贡献

大模型开源社区是一个开源项目的孵化和贡献平台。社区成员可以通过参与开源项目的开发和贡献来推动大模型技术的发展。通过分析社区成员参与开源项目的方式和贡献的内容，可以了解到社区成员对于开源项目的贡献和参与程度，以及他们如何通过开源项目来推动大模型技术的发展。

技术交流和分享

大模型开源社区是一个技术交流和分享的平台。社区成员可以通过技术演讲、技术文章、技术讨论等方式分享自己的经验和见解。通过分析社区成员的技术交流和分享方式，可以了解到社区成员之间的技术交流和学习方式，以及他们如何通过分享经验来促进大模型技术的发展。



社区治理和组织

大模型开源社区需要一定的治理和组织机制来保证社区的正常运行和发展。通过分析社区的治理和组织机制，可以了解到社区成员如何参与社区的决策和管理，以及他们如何通过社区的治理和组织机制来推动大模型技术的发展。

- 近年来，开源模型在人工智能领域迅速崛起，具有更好的透明度和可信赖性。尽管开源模型仍面临数据瓶颈和商业化的挑战，但**随着更多企业开源其模型，开源社区有望在数据建立和模型发展方面取得更大突破**。近期，Llama2项目引起广泛关注，开源社区参与者背景发生变化，商业公司也开始与开源社区合作。
- 在模型技术方面，大家关注点集中在模型大小、强度和商业化上。开源社区是开源项目从商业角度区别于其他商业模式的核心点，使潜在的免费用户变成社区的贡献者，产生价值。国内互联网大厂有技术能力，但语言壁垒限制了其在全球范围内的应用。**未来开源社区将解决语言壁垒，实现跨语言合作，国内社区将得到进一步发展。**

数据来源：甲子光年智库，2024年；

4.3 智源人工智能研究院大模型开源社区的典型经验分析

智源人工智能研究院（BAAI）

生态研发主体

是以其自身为核心，联合国内外的研究机构和企业共同推进的。智源研究院致力于构建以大模型为核心的生态系统，这不仅包括底层数据处理和汇聚、模型能力和算法评测，还包括开源开放的生态布局

运营模式

运营模式侧重于构建以大模型为核心的生态，这包括底层数据处理和汇聚、模型能力和算法评测、开源开放，形成一套高效的大模型技术和算法体系

北京智源 BAAI

投融资情况

- 北京智源人工智能研究是非盈利研发机构。
- 根据其非营利机构性质，智源研究院可能主要依赖于政府资助、科研项目经费和行业合作来支持其运营和研发活动。

开源生态布局

智源研究院推出了包括FlagAI、FlagPerf、FlagEval、FlagData、FlagBoot 和 FlagStudio 在内的FlagOpen（飞智）大模型技术开源体系，旨在支持多种深度学习框架和AI芯片，降低大模型开发的难度，助力全球开发者开展各种大模型的开发和研究工作

FlagEval
面向大规模基础模型的一体化评测平台

FlagVision
面向通用机器视觉的开源基础模型

FlagAI
集大模型算法和工具为一体的一站式开源大模型软件体系

FlagPerf
面向AI异构芯片的一体化基准性能评测引擎

FlagStudio
利用人工智能大模型支持艺术创作应用

FlagBoot
基于Scala开发的轻量级高并发微服务框架

FlagData
面向大模型研究领域的高效易用数据处理工具包

开源社区组织架构

- 智源研究院与多家产学研单位共同构建了大模型开源开放软件体系FlagOpen，这显示了其开源社区的合作性组织架构，旨在推动大模型软硬件生态的建设。
- FlagEval（天秤）大语言评测体系及开放平台是「科技创新 2030」旗舰项目重要课题，合作共建单位包括北大、北航、北师大、北邮、闽江学院、南开等高校和中科院自动化所、中国电子技术标准化研究院等科研院所，定期发布权威评测榜单。

重点应用领域

- 智源研究院的大模型技术主要应用于语言、视觉、多模态等基础大模型领域。
- 例如，"悟道·天鹰Aquila"语言大模型支持中英双语知识，"悟道·视界"视觉大模型系列解决了计算机视觉领域的一系列瓶颈问题

数据来源：BAAI，甲子光年智库，2024年；

4.4 百川智能大模型开源社区的典型经验分析

百川智能

生态研发主体

百川智能开源大模型生态的研发主体主要是百川智能公司。百川智能是一家专注于自然语言处理（NLP）和深度学习技术的创新型公司，拥有丰富的研发经验和专业技术团队，能够为大模型的研发和优化提供强大的技术支持。

运营模式

百川智能开源大模型生态采用开放、协作、共赢的运营模式。参与者可以通过开源社区共同开发和优化大模型，共享资源，互相学习，提高研发效率和应用效果。同时，百川智能还通过与合作伙伴、企业、科研机构等合作，共同打造各领域和行业的大模型，推动大模型的开源与应用。

百川智能 BAICHUAN AI

投融资情况

时间	轮次	融资额	投资方
2023年10月	A1轮	3亿美元	阿里巴巴、腾讯投资、小米集团、心资本、愉悦资本
2023年5月	天使轮	5000万美元	腾讯、小米、金山、慕华资本、清华大学资产管理有限公司、好未来、洪策资本、深创投、红点中国、卓源资本、众为资本、愉悦资本、顺为资本、心资本等十余家联合投资

开源生态布局

百川智能开源大模型生态已经建立了较为完善的开源生态布局。在模型库方面，百川智能已经开源了多个大模型，包括70亿参数量的Baichuan2-7B和130亿参数量的Baichuan2-13B等。这些模型的数据来自万亿互联网数据和垂直行业中的数据，并且训练的规模高达2.6TB。同时，百川智能还对模型训练进行了优化，使得在千卡A800集群中的训练性能达到了180TFLOPS，并且机器利用率超过50%。

Baichuan2-13B

Baichuan2-13B
开源可商用 低成本部署 多语言

130亿	2.6T	4K
模型参数量	多语言语料	大尺寸上下文

Baichuan2-7B

Baichuan2-7B
开源可商用 低成本部署 多语言

70亿	2.6T	4K
模型参数量	多语言语料	大尺寸上下文

开源社区组织架构

百川智能开源大模型生态的开源社区组织架构包括多个技术委员会、工作委员会和咨询委员会等。这些委员会由来自不同领域和行业的专家和开发者组成，负责技术决策、项目管理、社区运营等方面的工作。同时，百川智能还积极与合作伙伴、企业、科研机构等合作，共同打造各领域和行业的大模型，推动大模型的开源与应用。

重点应用领域

- 百川智能开源大模型生态的重点应用领域包括互联网、金融、医疗、教育等。这些领域对大模型的需求强烈，通过应用大模型可以提高效率、优化流程、改善用户体验等。
- 百川智能的大模型也对中、英、西、法等几十种语言提供支持，主要应用于学术研究、互联网和金融领域。

数据来源：百川智能，甲子光年智库，2024年；

目录

CONTENTS

甲子光年
JAZZYEAR

5.1 开源模型让每一家公司都具备成为AI公司的可能性

开源模型将覆盖更多企业和场景，具备创新自由度、用户体验等方面的优势

- 开源产品凭借更广泛的用户覆盖面和更大的创新自由度，在用户体验和技术创新方面具有明显优势，这是闭源产品难以企及的。
- 开源模型与闭源模型就像Linux与Windows，Android与iOS，互为竞争、互为补充。尽管闭源产品能更快、更直接地转化为商业利益，并因此加快产品迭代速度、提升服务质量，但开源模式所带来的用户粘性和技术创新动力仍是不可替代的。

开源产品优势

01 更广泛的用户覆盖面

- 大规模用户基础：更多用户参与，提供多样化的需求和反馈。
- 全球社区支持：来自世界各地的用户和开发者共同推动产品改进。

02 更大的创新自由度

- 无约束创新：开发者可自由探索和实现创新想法。
- 快速迭代：开放的交流和合作环境，加速技术进步和产品更新。

03 用户体验优势

- 用户驱动改进：用户反馈直接影响产品开发，提升用户满意度。
- 定制化能力：用户和企业可以根据自身需求定制和优化产品。

04 技术创新优势

- 社区智慧：集全球开发者智慧，推动技术前沿发展。
- 透明性和审查：开放代码便于审查和改进，确保高质量和创新性。

开源模型将激活每家企业，开源模型是让每一家公司都成为AI公司的关键因素。

数据来源：Data AI 2024，甲子光年智库，2024年；

5.2 开源大模型商业模式类型分析

开源模型是加快人工智能普及应用的关键

- 开源空间是一个边界封闭，内部开放的空间，受到现实世界和商业规则的约束。模型的开源会是保障大模型技术安全，解决安全漏洞的有效措施。
- 商业模式，与收入模型和成本结构有关。在开源方面，对外开源和使用开源所面对的商业模式有所不同。
- 从成本角度而言，对外开源所含的成本包括社区运营成本、开源安全成本；使用开源所含的成本包括开源合规成本、开源安全成本等；

开源大模型商业模式

模型开源，服务收费

- 模型开源免费使用；在定制开发、模型调优和部署等服务环节收费；
- 阿里云通义千问模型开源，通过云端部署大模型，以及一系列大模型部署解决方案收费；

01

通过公司其他业务来变现

- 该模式适合大型企业；该模式通过模型开源的免费方式吸引大量用户群体，在此基础上将通过成熟业务实现流量变现；
- Llama系列的开源让Meta保持在AI领域的独特优势，反哺社交、广告等业务的同时，也打开了和云服务商、AI硬件平台等合作的另一种可能。

02

通过生态来实现盈利

- 通过生态来实现盈利挑战性强，如Google的Android；Google的Play store是其最大的收入来源；
- Nvidia开源Nemotron-4 340B模型，它包含一系列开放模型，开发人员可以使用这些模型生成合成数据，用于训练大语言模型，可用于医疗健康、金融、制造、零售等所有行业。

03

开源获客，再推商业化

- 这类模式适合初创企业；
- 比如2023年成立的Mistral AI，因发布Mistral 7B、Mistral 8X7B两大开源模型一举成名，之后再推商业化模型，已能在业内快速打开局面；
- Mistral新推出的旗舰版大模型Large 则为闭源产品。

04

数据来源：甲子光年智库，2024年；

开源模型激活众多企业，应用于众多场景和领域

- 开源模型通过激活众多企业，广泛应用于各个领域和场景，推动了技术创新和行业发展，构建了一个充满活力和合作的生态系统。

开源模型应用众多场景，最终形成产业生态

广泛的企业应用

- **大公司**：加速研发进程，推动技术创新，开源产品使用；
- **中小企业**：降低技术门槛，使用开源模型，实现AI应用。
- **初创企业**：利用开源资源，通过切入细分场景，快速进入市场；

场景应用

- **金融**：欺诈检测、客户信用评估、自动化交易。
- **医疗**：疾病诊断、个性化治疗方案、药物研发。
- **零售**：个性化推荐、库存管理、客户行为分析。
- **制造**：预测性维护、质量控制、生产优化。

赋能创新企业

- **加速研发**：共享开源代码和模型，缩短开发周期。
- **降低成本**：减少专有软件费用，优化资源配置。
- **提升竞争力**：快速适应市场变化，推出创新产品和服务

构建开放生态系统

- **社区合作**：企业间合作，共同推动技术进步。
- **知识共享**：开放的知识和技术资源，提升整体行业水平。
- **标准化**：推动行业标准化，促进技术互操作性。

数据来源：甲子光年智库，2024年；

甲子光年
JAZZYEAR

谢 谢

北京甲子光年科技服务有限公司是一家科技智库，包含智库、媒体、社群、企业服务版块，立足于中国科技创新前沿阵地，动态跟踪头部科技企业发展和传统产业技术升级案例，致力于推动人工智能、大数据、物联网、云计算、AR/VR交互技术、信息安全、金融科技、大健康等科技创新在产业之中的应用与落地

关注甲子光年公众号

扫码联系商务合作

分析师

努尔麦麦提·买合木提（小麦）微信
13051317677

智库院长

宋涛微信
stgg_6406