头豹 LeadLeo

中国端侧大模型行业研究

算力优化与效率革命

如何重塑行业生态

企业标签：阿里云、商汤科技、面壁智能

AI变革行业创新发展

China End To Side Large Model Industry

中国エンド側大型モデル産業

撰写人：王利华

摘要

端侧大模型定义为运行在设备端的大规模人工智能模型，这些模型通常部署在本地设备上，如智能手机、IoT、PC、机器人等设备。与传统的云端大模型相比，端侧大模型的参数量更小，因此可以在设备端直接使用算力进行运行，无需依赖云端算力。

端侧大模型在成本、能耗、可靠性、隐私和个性化方面相比云端推理具有显著优势，并能够以低能耗提供高效且安全的AI处理，减少延迟并保护用户隐私，适合个性化的AI应用。取决于行业对数据安全、隐私保护的需求、行业本身智能设备的普及程度以及AI大模型技术的成熟度，这些因素的相互作用和共同推动，端侧大模型将推动各行业智能化发展的步伐。

端侧大模型面临的行业壁垒包括技术、硬件、数据、成本以及市场等方面，要求产业界在技术创新、标准制定、生态建设和市场推广等方面进行深入合作，以克服挑战，实现端侧大模型的广泛应用和落地。

- **2023年中国端侧大模型市场规模达8亿元，持乐观态度估计，预计2024年中国端侧大模型市场将达到21亿元**

生成式AI市场的蓬勃兴起，正驱使大模型厂商积极探索端侧应用新蓝海，以此作为增长的新引擎。端侧大模型通过在设备本地运行，有效降低了数据传输延迟，增强了隐私保护，拓宽了AI应用场景的广度与深度。

与此同时，下游市场需求的强劲增长，特别是手机与自动驾驶行业的蓬勃发展，正强力拉动端侧大模型市场的扩张，2023年中国端侧大模型市场规模达8亿元，预计2024年中国端侧大模型市场将达到21亿元。

依托技术实力和生态建设，头部大模型厂商纷纷投入端侧大模型市场，利用在云端大模型领域的技术优势，商汤商量、阿里通义以及面壁智能率先在端侧大模型领域取得领先突破。

研究框架

名词解释

- **AI大模型：**指的是大型人工智能模型，通常由数十亿至数百亿个参数组成，用于各种自然语言处理、计算机视觉等任务。
- **模型压缩技术：**是一系列用于减少大型神经网络模型尺寸和计算复杂度的技术，包括剪枝、量化、蒸馏等方法，旨在减少模型大小的同时保持其性能。
- **IoT设备：**指的是物联网设备，通常具有较小的计算能力和存储空间，但能够通过互联网进行通信和数据交换。
- **PC设备：**个人计算机，如台式机、笔记本电脑等，通常具有较高的计算和存储能力，适合运行复杂的应用程序和任务。
- **数据中心：**指的是大规模的服务器集群，用于存储和处理大量数据，支持云计算服务和网络应用。
- **服务器：**通常指的是提供网络服务、存储和计算资源的计算机系统，可用于托管网站、应用程序等。
- **BERT：**是一种预训练的自然语言处理模型，采用Transformer架构，能够理解文本语境并在各种NLP任务中取得良好性能。
- **DistilBERT：**是对BERT模型进行了蒸馏（Distillation）的轻量化版本，通过减少参数和计算复杂度来提高模型的运行效率。
- **TinyBERT：**是进一步轻量化的BERT模型，通过更深入的模型压缩和优化来适应资源受限的环境，如移动设备或物联网设备。
- **Jetson AGX Xavier：**高性能嵌入式系统，具有GPU和AI计算能力，适用于边缘计算和深度学习应用。
- **TPU：**谷歌推出的张量处理单元，是一种专门用于加速人工智能工作负载的定制硬件加速器。
- **PyTorch Mobile：**是PyTorch框架的移动端部署版本，支持在移动设备上运行训练好的深度学习模型。
- **TensorFlow Lite：**是谷歌推出的用于在移动设备和嵌入式系统上部署深度学习模型的轻量级框架。
- **ONNX：**开放神经网络交换，是一种开放的跨平台深度学习模型表示格式，支持模型在不同框架之间的转换和部署。
- **预训练模型：**指的是在大规模文本数据上进行预训练的神经网络模型，通常包含通用的语言或视觉理解能力，并可通过微调适应特定任务。
- **中心云：**指的是传统的云计算架构，数据和计算资源集中在大型数据中心进行管理和运行。
- **边缘云：**是一种分布式的云计算架构，将计算和存储资源放置在接近终端用户的边缘节点上，以提高服务响应速度和降低网络延迟。
- **AI芯片：**专门用于加速人工智能计算任务的硬件芯片，能够在高效率和低能耗的条件下进行大规模并行计算。
- **知识蒸馏：**是一种通过让一个较大且性能较好的模型（教师模型）指导一个小型模型（学生模型）来提高学生模型性能的技术，通常用于模型压缩和轻量化。

Chapter 1
行业概述

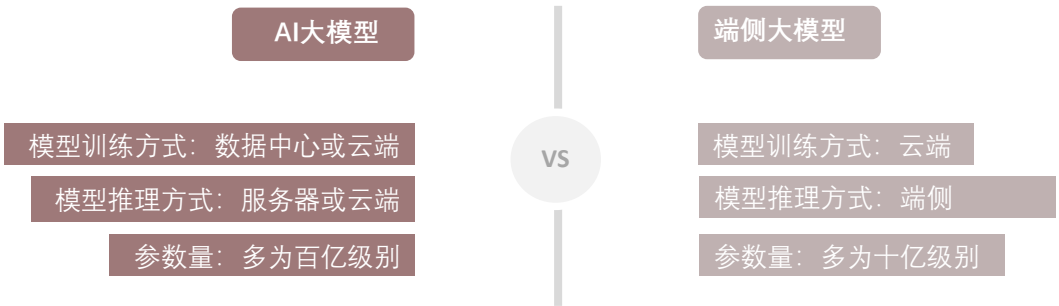
- 定义与分类
- 发展历程
- 驱动力
- 市场规模

中国端侧大模型市场探析——定义与分类

- 端侧大模型定义为运行在设备端的大规模人工智能模型，与传统的云端大模型相比，端侧大模型的参数量更小，因此可以在设备端直接使用算力进行运行，无需依赖云端算力

端侧大模型的定义

- 端侧大模型定义为运行在设备端的大规模人工智能模型，这些模型通常部署在本地设备上，如智能手机、IoT、PC、机器人等设备。与传统的云端大模型相比，端侧大模型的参数量更小，因此可以在设备端直接使用算力进行运行，无需依赖云端算力。



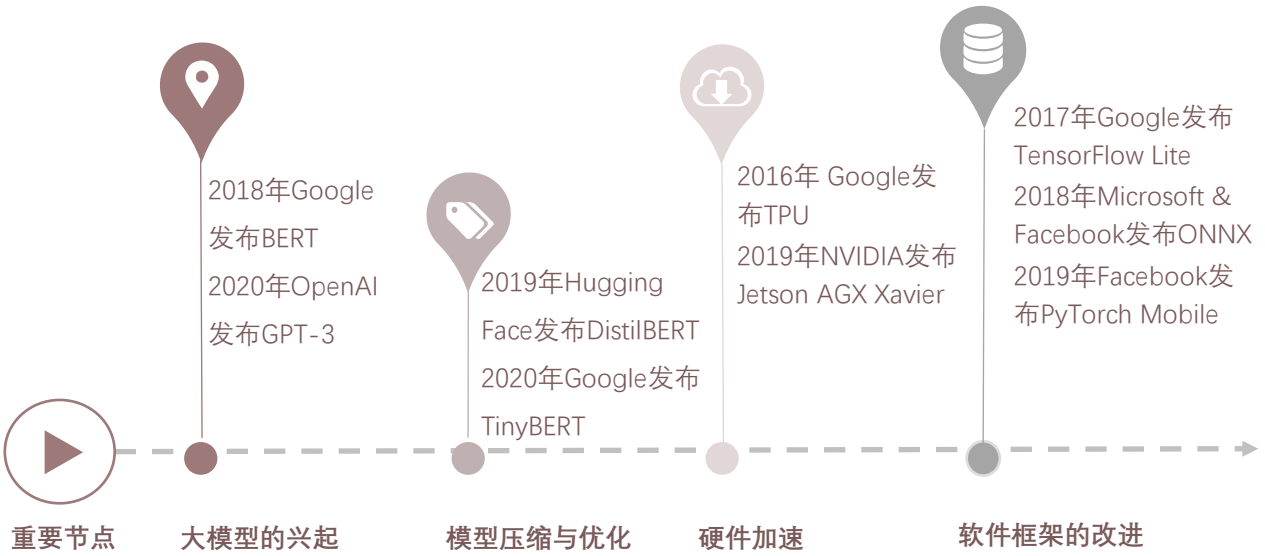
- AI大模型通常在数据中心或云端进行训练，使用大规模的计算资源和海量数据。相比之下，端侧大模型由于资源限制，往往需要在设计和训练阶段进行模型压缩和优化。在推理方式上，AI大模型通常运行在服务器或云端，通过强大的计算能力处理复杂的任务。然而，这种云端推理方式依赖于网络连接，会带来延迟和隐私问题。端侧大模型则是在本地设备上进行推理。
- 参数量是AI大模型和端侧大模型的一个显著区别。AI大模型通常具有数十亿甚至上百亿的参数，如GPT-3的1,750亿参数。这种巨大的参数量使得大模型能够捕捉复杂的数据模式并在多种任务中表现出色。然而，端侧设备的计算能力和存储资源有限，因此端侧大模型的参数量通常较小。通过模型压缩技术，如知识蒸馏、剪枝和量化，端侧大模型的参数量可以减少到几百万或更少。例如，MobileBERT的参数量仅为BERT的1/4左右，但依然能够在移动设备上高效运行。

来源：企业官网，头豹研究院

中国端侧大模型市场探析——发展历程

- AI大模型逐渐在技术、硬件和应用层面实现向端侧设备的迁移和优化，提升端侧大模型在实际应用中的效率和性能，逐渐能够在端侧设备上高效运行，提供更好的用户体验和更多的实时应用场景

端侧大模型行业的发展历程分析



大模型的兴起

2018年，Google发布了BERT，这是第一个使用双向Transformer的预训练模型，在多个自然语言处理任务上取得了显著的性能提升。

2020年，OpenAI发布了GPT-3，它具有1750亿参数，展示了大规模语言模型在各种应用中的强大能力，并进一步推动了大模型的发展。

模型压缩与优化

通过知识蒸馏技术，将大模型的知识转移到小模型中，使得模型在保持较高性能的同时，减少计算资源需求。例如，TinyBERT和DistilBERT都是通过蒸馏技术获得的小型化模型。

硬件加速

Google的TPU和其他厂商的NPU专门用于加速AI模型的训练和推理，大大提升了大模型在设备端的性能。NVIDIA Jetson、华为Ascend等，提供了强大的边缘计算能力，使得大模型能够在终端设备上高效运行。

软件框架的改进

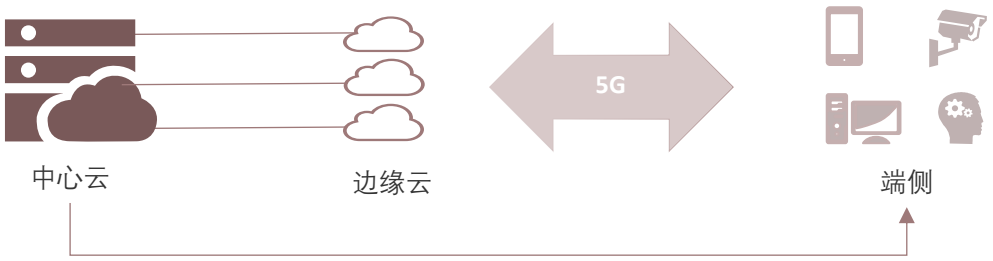
TensorFlow Lite和ONNX这些轻量级的模型推理框架支持在移动设备和嵌入式设备上运行深度学习模型，优化了资源利用和运行效率。PyTorch的移动版本，使得开发者能够更容易部署PyTorch模型。

来源：专家访谈，头豹研究院

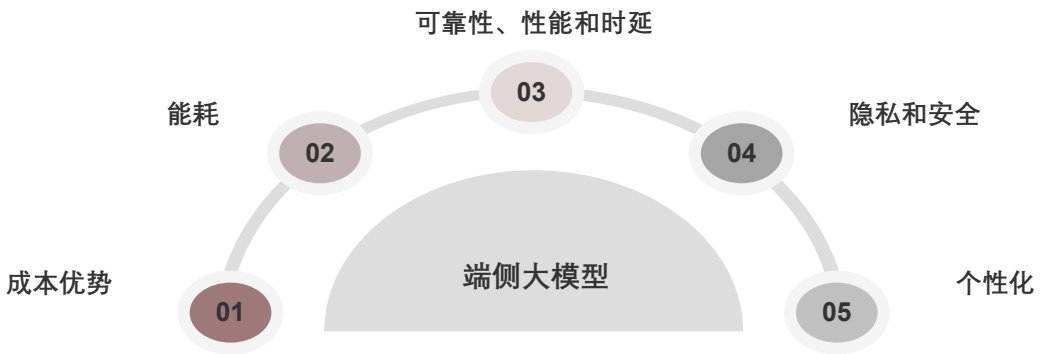
中国端侧大模型市场探析——驱动力

- 端侧大模型在成本、能耗、可靠性、隐私和个性化方面相比云端推理具有显著优势，并能够以低能耗提供高效且安全的AI处理，减少延迟并保护用户隐私，适合个性化的AI应用

端侧大模型市场驱动力分析



为实现规模化扩展，AI处理的重心，正在向边缘转移



- 从成本优势来看，AI推理的规模远高于AI训练。尽管训练单个模型会消耗大量资源，但大型生成式AI模型预计每年仅需训练几次。然而，这些模型的推理成本将随着日活用户数量及其使用频率的增加而增加。在云端进行推理的成本极高，这将导致规模化扩展难以持续。
- 从能耗来看，支持高效AI处理的边缘终端能够提供领先的能效，尤其是与云端相比。边缘终端能够以很低的能耗运行生成式AI模型，尤其是将处理和数据传输相结合时。这一能耗成本差异非常明显。
- 从可靠性、性能和时延来看，终端侧AI处理能够在云服务器和网络连接拥堵时，提供媲美云端甚至更佳的性能 。当生成式AI查询对于云的需求达到高峰期时，会产生大量排队等待和高时延，甚至出现拒绝服务的情况。向边缘终端转移计算负载可防止这一现象发生。
- 从隐私和安全来看，端侧大模型从本质上有助于保护用户隐私，因为查询和个人信息完全保留在终端上。对于企业和工作场所等场景中使用的生成式AI，这有助于解决保护公司保密信息的难题。
- 从个性化来看，数字助手将能够在不牺牲隐私的情况下，根据用户的表情、喜好和个性进行定制。所形成的用户画像能够从实际行为、价值观、痛点、需求、顾虑和问题等方面来体现一个用户，并且可以随着时间推移进行学习和演进。

来源：中国统计局，CNNIC，头豹研究院

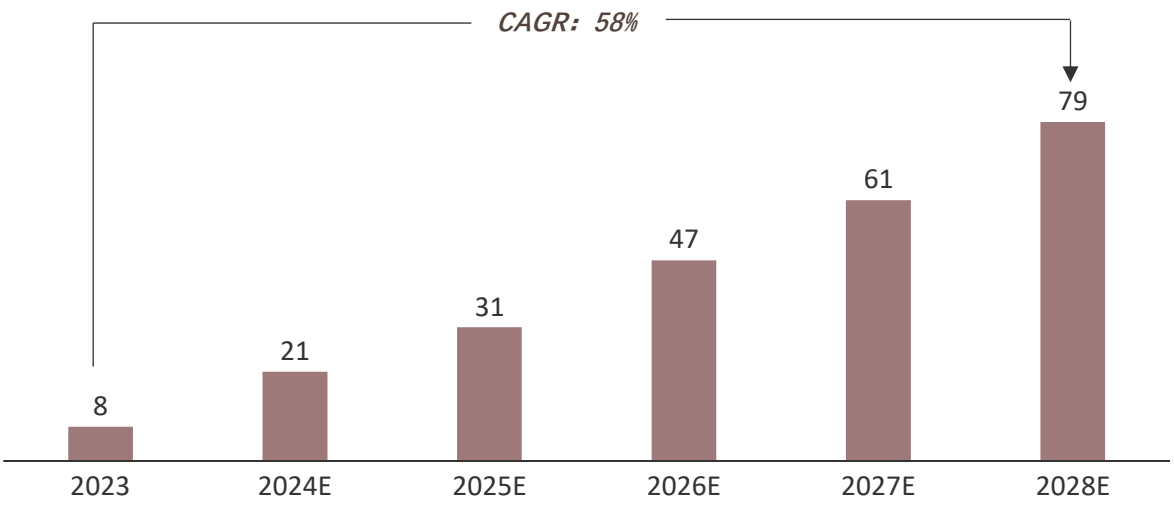
中国端侧大模型市场探析——市场规模

- 下游市场需求的强劲增长，特别是手机与自动驾驶行业的蓬勃发展，正强力拉动端侧大模型市场的扩张，2023年中国端侧大模型市场规模达8亿元，预计2024年中国端侧大模型市场将达到21亿元

中国端侧大模型行业——市场规模

中国端侧大模型市场规模

单位：亿元



- **受实际落地情况的影响，2023年中国端侧大模型市场规模达8亿元，持乐观态度估计，预计2024年中国端侧大模型市场将达到21亿元**

生成式AI市场的蓬勃兴起，正驱使大模型厂商积极探索端侧应用新蓝海，以此作为增长的新引擎。端侧大模型通过在设备本地运行，有效降低了数据传输延迟，增强了隐私保护，拓宽了AI应用场景的广度与深度。例如，智能手机集成的AI摄影功能，能实时识别场景并优化图像质量；可穿戴设备利用端侧模型监测健康指标，提供即时反馈。与此同时，随着AI芯片等算力市场带动，为端侧大模型打开新的市场空间。高性能、低功耗的AI芯片设计使得复杂模型能够在手机、物联网设备等终端高效运行，无需依赖云服务，显著提升响应速度与用户体验。2021年全球AI芯片市场规模达到200亿美元，预计到2025年将超过700亿美元，其中端侧AI芯片占比快速提升，成为增长的重要动力。

- **下游市场需求的强劲增长，特别是手机与自动驾驶行业的蓬勃发展，正强力拉动端侧大模型市场的扩张**

手机作为个人智能终端的核心，正集成更先进的AI功能以提供个性化服务与优化用户体验，如荣耀Magic系列利用端侧AI大模型实现偏好理解与多模态交互。同时，自动驾驶领域对实时性与安全性要求极高，推动了BEV+Transformer等技术与端侧大模型的融合，百度Apollo ADFM等L4级自动驾驶大模型的推出，标志着该领域迈向商用新阶段。

来源：专家访谈，企业公告，头豹研究院

Chapter 2
产业链分析

- 产业链图谱
- 模型压缩技术
- 成本构成
- 厂商类型
- 行业场景
- 业务场景

中国端侧大模型市场探析——产业链

- 中国端侧大模型上游主要包括AI芯片供应商、云计算服务商以及数据服务商，中游为端侧大模型科技厂商和端侧科技企业，主要通过设备企业最终应用到汽车、教育等各行各业

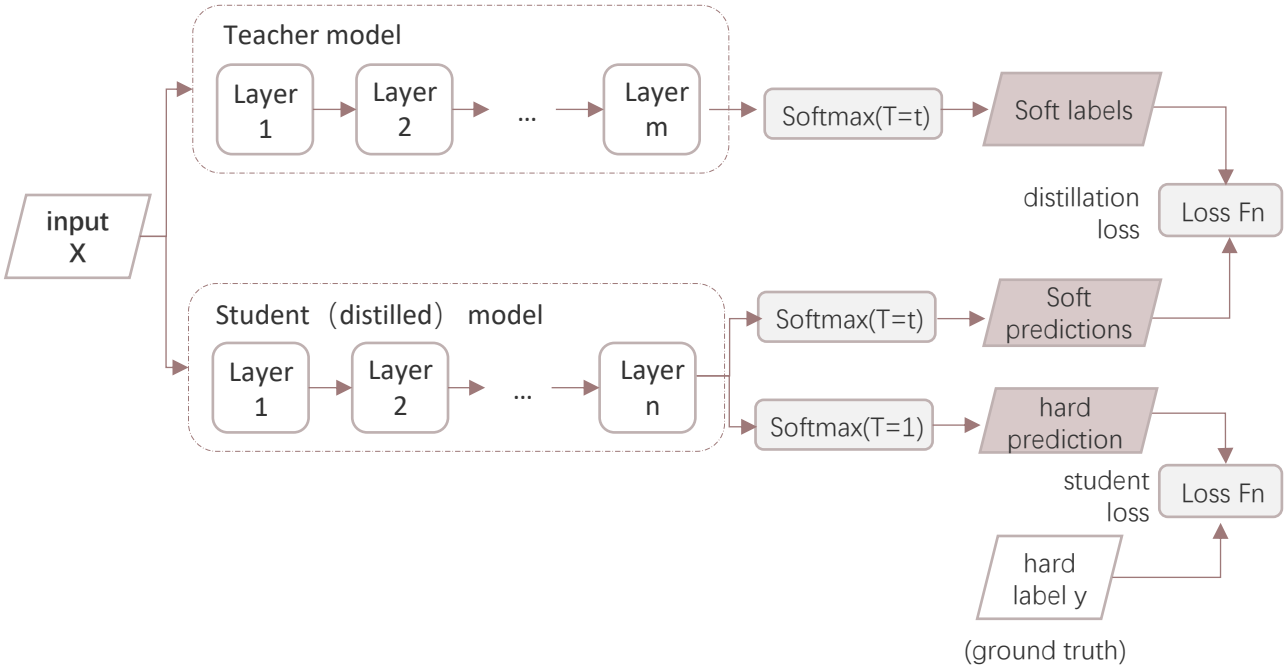
中国端侧大模型行业——产业链分析

来源：专家访谈，头豹研究院

中国端侧大模型市场探析——模型压缩技术

- 通过知识蒸馏，端侧大模型能够在保持较高性能的同时，大幅减少模型参数量和计算复杂度。这种技术使得复杂的AI模型可在资源受限的端侧设备上高效运行，实现低能耗、高响应速度和高准确度的AI推理

端侧大模型压缩技术——知识蒸馏



■ **知识蒸馏的基本原理**

首先，在强大的计算资源和海量数据集上训练一个高性能的大模型，称为教师模型。教师模型在输入训练数据时，不仅输出最终的分类结果（硬标签），还输出每个类别的概率分布（软标签），这些软标签包含了更多关于输入数据的细微信息和模式。在训练较小的学生模型时，不仅使用原始数据的硬标签，还使用教师模型生成的软标签。学生模型通过学习这些软标签，能够捕捉到教师模型中包含的丰富知识。

■ **知识蒸馏在端侧大模型中的应用**

知识蒸馏使得学生模型能够在保持较高准确度的同时，显著减少参数量。例如，TinyBERT通过知识蒸馏技术将BERT的参数量减少到原来的1/7左右，但在许多自然语言处理任务中仍能保持较好的性能。这使得学生模型能够适应端侧设备的计算和存储限制。较小的学生模型在推理阶段需要的计算资源更少，推理速度更快。这对于资源受限的端侧设备尤为重要。

端侧设备通常对能耗有严格限制。知识蒸馏生成的学生模型由于计算复杂度低，能够以较低的能耗完成推理任务。例如，在物联网设备和移动设备中，学生模型的低能耗运行方式使其能够长时间持续工作，而不会显著消耗电池电量。

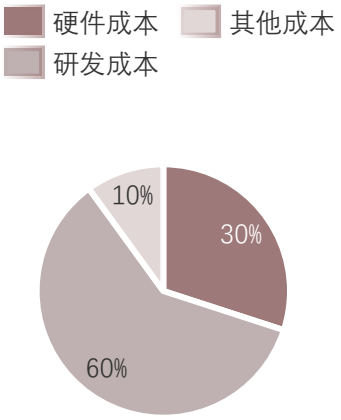
知识蒸馏生成的学生模型可以针对不同的端侧设备进行优化。例如，针对特定硬件架构进行剪枝和量化，使模型在特定设备上达到最佳性能。此外，学生模型还可以通过在线学习机制，在端侧设备上不断适应和优化，以满足动态变化的应用需求。

来源：专家访谈，智慧文旅，头豹研究院

中国端侧大模型市场探析——成本构成

- AI芯片作为加速端侧大模型的关键技术，提供高效计算和能耗比，使得大规模模型在端侧设备上高效运行，研发人员及显卡成本需兼顾，确保研发经济可持续

端侧大模型成本构成分析



硬件成本包括端侧大模型部署所必需的硬件设备，如AI芯片。AI芯片专门设计用于加速深度学习任务，相比通用处理器，AI芯片可提供更高的计算效率，降低模型执行的能耗和延迟。

研发成本是开发和优化端侧大模型所需的成本，主要为研发人员的人力成本以及设备成本。人力成本涉及到研究人员、工程师和数据科学家等的工资、培训和福利待遇。设备成本包括用于开发和测试的显卡、服务器和云服务等。

其他成本包括除了硬件和研发成本之外，还包括各种间接成本，例如管理成本、运营成本和市场推广成本等。

AI芯片成为加速端侧大模型应用的关键技术成本

AI芯片作为专门设计用于加速深度学习任务的硬件，具有较高的能效比和计算性能，成为了实现端侧大模型高效部署的关键。一方面，AI芯片能够提供更高的计算性能和能效比，从而加速端侧大模型的推理和执行速度。例如，Google的TPU能够在相同的功耗下实现比传统GPU更高的性能，这使得在端侧设备上运行大规模的神经网络模型成为可能。另一方面，AI芯片也能够提供更低的功耗和更小的尺寸，适合嵌入到各种端侧设备中，为端侧大模型的应用提供了更广泛的可能性和更好的用户体验。

在端侧大模型的开发过程中，需要综合考虑研发人员的成本和显卡的成本，以确保项目的顺利进行和成功实施

深度学习模型的研发需要具有深度学习和机器学习背景的专业人员，他们负责模型的设计、算法优化、超参数调整等工作。美国的机器学习工程师的平均年薪约为12万美元，而深度学习工程师的平均年薪则更高，约为14万美元。因此，合理控制研发人员的成本，并保证其具备高水平的技能和专业知识，对于端侧大模型的研发和应用至关重要。其次，显卡的性能和规模直接影响着模型训练的速度和效率。一台高端显卡如NVIDIA GeForce RTX 3,090的价格约为1,500美元。此外，显卡可以通过云服务提供商租用，这也是许多企业在进行端侧大模型的研发和优化时采取的一种常见方式。但在长期使用过程中，租用成本也会成为企业的一项不小的支出。因此，企业需要综合考虑研发人员的成本、显卡租用成本以及其他相关成本，以确保研发过程的经济性和可持续性。

来源：专家访谈，智慧文旅，头豹研究院

中国端侧大模型市场探析——厂商类型

- 端侧大模型厂商通过许可费、订阅和定制化服务模式为企业客户、互联网公司和开发者提供智能化解决方案；而终端设备厂商则以硬件销售和增值服务为主，为企业和消费者提供集成端侧大模型的智能设备

中国端侧大模型行业——商业模式分析

	端侧大模型厂商	终端设备厂商
主要业务	开发、训练和优化端侧大模型	将端侧大模型集成到其产品中，如AI手机、AI PC等
客户群体	**各行业企业：** 包括各行各业的企业，如金融、医疗、零售等，需要使用端侧大模型来优化业务流程、改善用户体验等。	**普通消费者：** 即最终使用终端设备的个人用户，他们有需求购买智能手机、智能音箱、智能电视等产品。
	终端设备厂商: 包括个人开发者和软件开发公司，需要端侧大模型来集成到自己的产品中，提供更智能的功能和服务。	**企业客户：** 企业需要大量购买终端设备，用于员工办公、生产管理、客户服务等方面。
商业模式	许可费模式、订阅模式、定制化开发等	硬件销售模式、增值服务模式等

- **端侧大模型厂商主要通过提供端侧大模型服务来获取收入，其商业模式包括许可费模式、订阅模式和定制化开发等**

端侧大模型厂商的服务模式可以面向企业客户、互联网公司以及应用开发者等不同的客户群体。企业客户会购买端侧大模型服务，用于优化业务流程、改善用户体验等方面。互联网公司会订阅端侧大模型服务，以改善推荐系统、广告投放等功能。而应用开发者则会定制化开发端侧大模型，集成到自己的应用程序中，提供更智能化的功能和服务。

- **终端设备厂商主要通过硬件销售模式和增值服务模式获利**

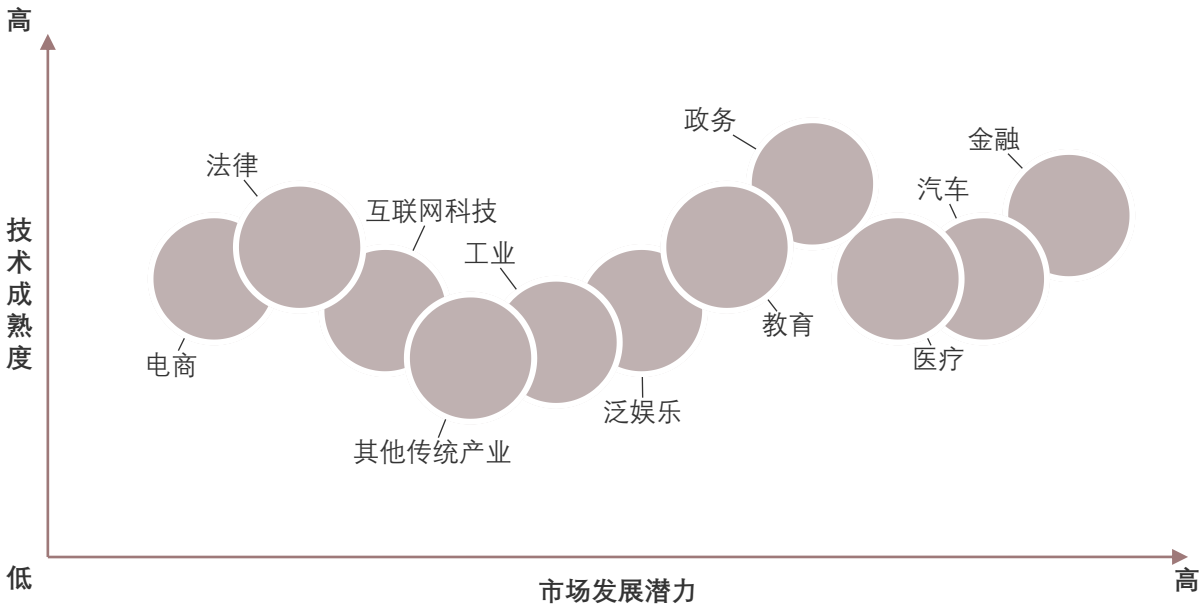
终端设备厂商将端侧大模型集成到智能手机、智能音箱等产品中，并通过销售这些产品来获取收入。同时，他们也为企业客户提供定制化服务，满足其员工办公和生产管理等需求；针对普通消费者，终端设备厂商提供智能化产品，如智能手机、智能音箱，享受其中提供的端侧大模型服务；此外，开发者和合作伙伴也与终端设备厂商合作，开发定制化应用程序或服务，利用端侧大模型实现更丰富的功能和体验。

来源：专家访谈，头豹研究院

中国端侧大模型市场探析——行业场景

- 取决于行业对数据安全、隐私保护的需求、行业本身智能设备的普及程度以及AI大模型技术的成熟度，这些因素的相互作用和共同推动，端侧大模型将推动各行业智能化发展的步伐

中国端侧大模型行业——行业场景分析



行业对数据安全和隐私保护的需求将直接影响端侧大模型的应用

随着数据泄露和隐私问题的日益突出，各行业对于数据的保护需求愈发迫切。因此，在端侧大模型的应用中，需要采取一系列的技术手段来确保数据的安全性和隐私性，如联合学习、加密计算等。这将促使行业在应用端侧大模型时更加谨慎和审慎，但也为解决数据安全隐患提供了新的解决方案。因此，端侧大模型在金融、医疗、政务等对数据安全要求较高的行业具有较大发展潜力。

行业本身智能设备的普及程度也将影响端侧大模型的发展前景

随着智能设备的普及程度提高，对于端侧AI应用的需求也将相应增加。这些智能设备不仅提供了丰富的数据来源，也为端侧大模型的运行提供了更多的计算资源和场景。例如，随着智慧教堂的普及率加深，教育成为端侧大模型未来发展的潜力场景之一。此外，在医疗领域，家用健康监测设备能够使数据存储在设备端，更能满足客户的隐私性。

AI大模型技术的成熟度是端侧大模型发展的重要因素之一

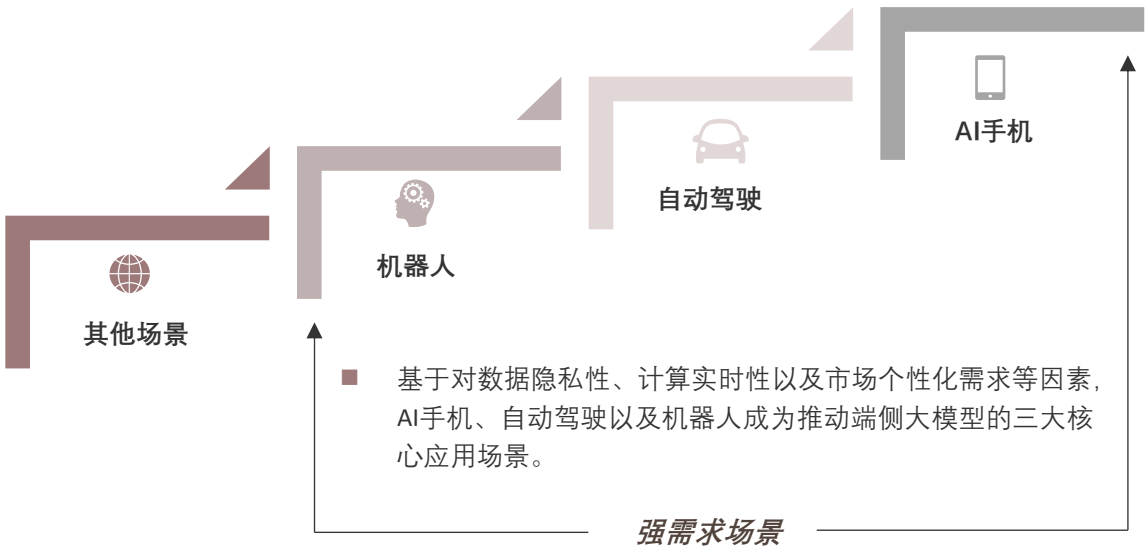
端侧大模型应用依赖于AI大模型的技术基础，随着AI大模型在自然语言处理、计算机视觉、语音识别等领域的发展和成熟，端侧大模型应用也得到推动；各行业对端侧设备上运行的大型AI模型的需求不断增加，促使端侧大模型应用成熟度与AI大模型保持一致；同时，技术转移和跨界应用使得一些在特定行业中成熟的AI大模型技术可以被应用到其他行业的端侧设备中，推动两者的同步发展。

来源：专家访谈，头豹研究院

中国端侧大模型市场探析——业务场景

- 端侧大模型能在保障数据隐私的同时，实现低延迟的实时计算，并提供高度个性化的服务，因此基于对数据隐私、计算实时以及个性化等强需求，AI手机、自动驾驶和机器人成为端侧大模型核心应用场景

端侧大模型业务场景分析——按不同的设备类型分类



AI手机：数据隐私性和计算实时性

现代智能手机在语音助手、图像识别、自然语言处理等方面广泛应用AI技术，这些功能需要处理大量的用户数据。如果将这些数据传输到云端进行处理，不仅增加了隐私泄露的风险，还会由于网络延迟导致用户体验下降。端侧大模型可以在本地设备上进行数据处理，确保用户的敏感信息不被泄露，同时大幅提升计算的实时性。

自动驾驶：实时决策和安全性

自动驾驶车辆需要在复杂的道路环境中实时做出决策，如避让行人、识别交通信号和处理突发情况。这些决策需要极低的延迟，因为任何延误都可能带来安全隐患。端侧大模型能够在车辆本地进行高效计算，确保实时响应和高精度决策，从而提升自动驾驶系统的安全性和可靠性。此外，端侧计算减少了对网络连接的依赖，在网络条件不佳的情况下仍能保持车辆的正常运行。

机器人：个性化服务和效率提升

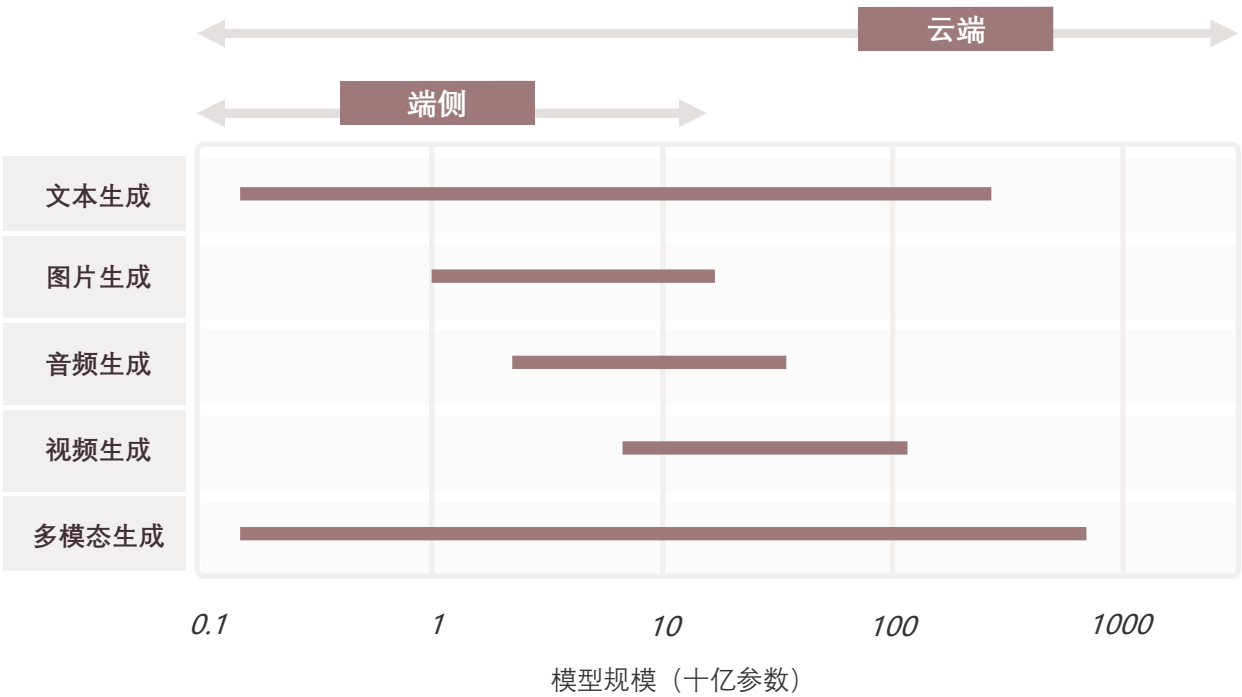
在机器人应用中，特别是家庭服务机器人和医疗机器人，端侧大模型的优势在于能够提供个性化服务和提高效率。机器人需要根据用户的具体需求和偏好调整其行为和功能，例如在家庭中提供特定的照护服务或在医院中执行特定的医疗任务。通过在本地处理数据，机器人可以更好地理解用户的行为和习惯，从而提供更加个性化的服务。此外，端侧计算能够提升机器人的响应速度和任务执行效率，在面对复杂环境和任务时，机器人可以迅速做出决策和调整，确保任务的准确性和高效性。这种能力不仅提高了用户体验，还拓展了机器人的应用范围和市场潜力。

来源：专家访谈，头豹研究院

业务场景（接上页）

- 随着技术的不断进步和应用场景的拓展，端侧大模型各业务场景中存在差异，文本生成和图片生成场景相对较成熟，音频生成场景逐步发展，视频生成和多模态生成场景尚处于起步阶段

端侧大模型业务场景分析——按不同的技术场景分类



■ **文本生成与图片生成的业务场景**

文本生成模型如GPT系列在端侧的应用逐渐成熟，尤其是在智能手机等移动设备上的应用。通过模型压缩和优化，现有的文本生成模型已经可以在资源受限的环境下高效运行。图片生成模型的端侧应用也在逐步发展，尤其是一些轻量级的图像生成模型。这些模型可以用于图像风格转换、图像修复、图像增强等应用，为用户提供更丰富的图像处理功能。随着硬件技术的进步和模型算法的改进，图片生成模型在端侧的应用将进一步成熟。

■ **音频生成的业务场景**

音频生成模型在端侧的应用相对较新，但也在不断发展。目前一些语音合成模型已经可以在端侧设备上实现实时的语音合成功能，如智能语音助手、语音提示等。

■ **视频生成和多模态生成的业务场景**

相比于文本和图片生成模型，视频生成模型的端侧应用相对较少，主要原因是视频数据的复杂性和处理量较大。而一些视频压缩和编解码技术的进步以及硬件加速器的应用，为视频生成模型在端侧的应用提供一定的可能性。多模态生成模型是指同时处理多种类型数据的生成模型，其在端侧的应用也在逐步探索和发展，如智能多模态搜索、多模态推荐系统等，但其成熟度相对较低，需要更多的研究和技术突破。

来源：专家访谈，头豹研究院

Chapter 3
行业分析

- 政策分析
- 行业壁垒
- 竞争格局
- 发展趋势

中国端侧大模型市场探析——政策分析

- 中国政府将人工智能产业视为中国国家战略核心，在端侧大模型方面展现出积极的支持立场。在AI基础设施以及生成式AI方面设立规范，整体政策环境对AI产业及端侧大模型的健康发展表现有利

中国端侧大模型相关政策，2020-2024年

政策名称	颁布日期	颁布主体	主要内容及影响
《针对生成式人工智能服务出台管理办法》	2023-04	网信部	一方面，该办法支持人工智能算法、框架等基础技术的自主创新、推广应用、国际合作，为端侧大模型发展提供了政策支持和技术保障；另一方面，该办法要求端侧大模型在数据来源、算法设计、内容标识等方面遵守法律法规的要求，尊重社会公德、公序良俗，防止生成虚假信息、侵犯他人权益、造成社会不良影响等问题，为端侧大模型发展提供了规范引导和监督约束
《数字中国建设整体布局规划》	2023-02	国务院	不仅在技术基础、数据资源、应用场景、技术创新和政策环境等多个层面提供了支持和指导，还明确了发展方向和合规要求，为端侧大模型的健康、快速发展铺平了道路。这促使相关企业需不断提升技术创新能力，加强数据安全与隐私保护，深化与实体经济的融合，以适应并推动数字中国建设的总体布局
《关于加快场景创新以人工智能高水平应用促进经济高质量发展的指导意见》	2022-07	科技部	一方面，该指导意见鼓励在各行业领域深入挖掘人工智能技术应用场景，为端侧大模型提供了丰富多样的应用场景，如聊天和文本生成、机器翻译、语音识别与合成、自然语言理解与推理等；另一方面，该指导意见强调以需求为牵引谋划人工智能技术应用场景，为端侧大模型提供了需求驱动的动力，促进端侧大模型在解决实际问题中优化升级
《关于促进新一代人工智能产业高质量发展的若干措施》	2022-01	教育部	发挥科技支撑和引领作用，支持有条件的地区和高校、科研机构、企业开展语言智能技术研究，着力在自然语言处理、机器写作、机器翻译、机器评测等领域取得实质成果，为端侧大模型奠定技术实力
《工业和信息化部关于开展信息通信服务感知提升行动的通知》	2021-11	工信部	从事互联网信息服务的企业应建立客服热线电话，并在网站、APP等显著位置公示客服热线电话号码。鼓励具备条件的企业提供充足的人工客服坐席
《国家新一代人工智能标准体系建设指南》	2020-07	网信办	指南规划了新一代人工智能标准体系的总体框架和具体内容，包括标准目录、标准体系结构、标准分类和标准制定程序等。通过建设完备、系统、规范的人工智能标准体系，促进人工智能技术的创新和应用，保障人工智能的安全和可持续发展

来源：政府各部门，头豹研究院

中国端侧大模型市场探析——行业壁垒

- 端侧大模型面临的行业壁垒包括技术、硬件、数据、成本以及市场等方面，要求产业界在技术创新、标准制定、生态建设和市场推广等方面进行深入合作，以克服挑战，实现端侧大模型的广泛应用和落地

中国端侧大模型行业——行业壁垒分析

技术复杂性与硬件限制

实现高效的端侧大模型需要复杂的模型压缩和优化技术，需要深入的专业知识和丰富的实践经验，且不同模型和应用场景需要定制化的优化策略，增加了开发难度。

端侧设备的计算资源和存储空间通常有限，这对模型的尺寸和计算效率提出了严格的要求，限制了大模型在端侧设备上的应用。同时，端侧设备通常依赖电池供电，能耗管理非常重要，这需要模型在设计和运行时充分考虑能耗优化。

数据隐私和安全

在端侧设备上处理敏感数据需要严格的隐私保护措施。确保数据在本地处理过程中不被泄露或滥用是一个重要的挑战，需要在模型设计和应用过程中集成强大的数据加密和隐私保护机制。

开发和维护成本

开发和优化端侧大模型需要投入大量的资源，包括高性能计算设备、专业技术人员以及长期的研发投入。这对于中小型企业和初创公司来说，会构成较大的经济压力。

行业标准及市场应用

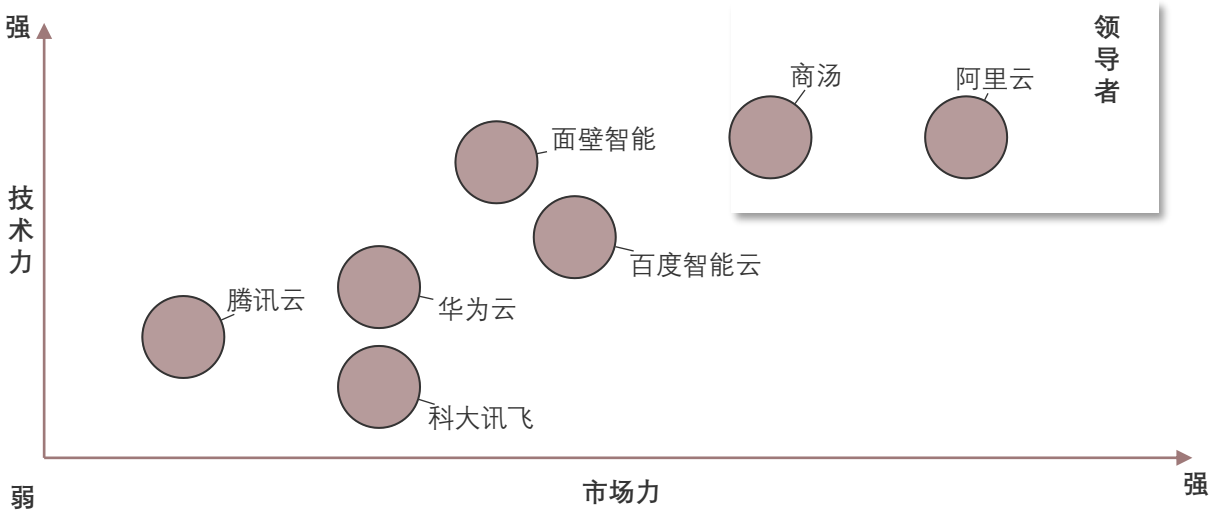
端侧大模型在不同的应用场景和设备平台上缺乏统一的标准和规范，这容易导致开发过程中出现兼容性和互操作性问题。行业需要制定和推广相关标准，促进端侧大模型的广泛应用。此外，尽管端侧大模型具有许多优势，但在某些应用场景下，市场需求尚未完全显现或开发出来。这需要行业参与者进一步挖掘和培育新的应用场景，推动技术的商业化落地。并且端侧大模型在实际应用中的性能、可靠性和用户体验需要得到用户的认可和信任，这需要时间和实际应用的验证。

来源：企业官网，头豹研究院

中国端侧大模型市场探析——竞争格局

- 依托技术实力和生态建设，头部大模型厂商纷纷投入端侧大模型市场，利用在云端大模型领域的技术优势，商汤商量、阿里通义以及面壁智能率先在端侧大模型领域取得领先突破

中国端侧大模型行业——竞争格局



依托技术实力和生态建设，头部大模型厂商纷纷投入端侧大模型市场

头部大模型厂商依托其深厚的技术积累和成熟的生态系统，正加速布局端侧大模型市场。一方面，这些厂商利用在云端大模型领域的技术优势，通过算法优化、模型压缩等先进技术，有效解决了端侧算力限制问题，使得复杂的AI功能能够在移动设备、物联网终端等平台上高效运行，满足用户对即时性、隐私保护及离线使用的需求，如商汤发布1.8B端侧大模型，阿里也发布18亿参数的通义端侧大模型。另一方面，通过构建开放的生态平台，整合上下游资源，赋能开发者与行业伙伴，共同探索端侧AI的多元化应用场景。

技术融合与创新驱动将加剧端侧大模型市场竞争

随着端侧大模型技术的日益成熟，未来中国端侧大模型行业的竞争格局将呈现出技术深度融合与创新驱动的新态势。一方面，技术融合将成为竞争的核心要素。厂商不再局限于单一技术的优化，而是趋向于跨领域技术的集成，如将自然语言处理、计算机视觉、边缘计算等技术与大模型结合，打造综合型AI解决方案。

生态系统构建与合作模式的创新将成为塑造竞争格局的关键

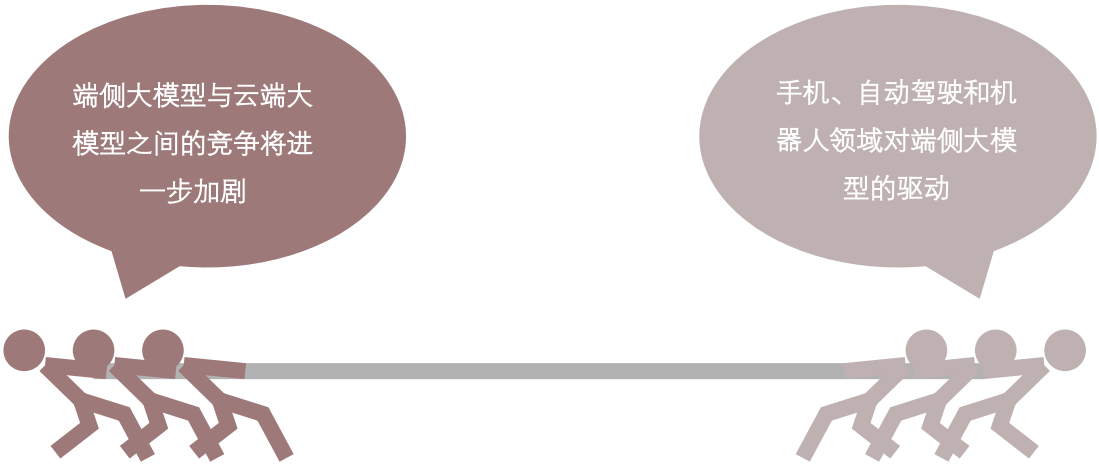
在端侧大模型的部署与应用中，单一企业的力量难以覆盖全部产业链环节，因此构建开放合作的生态系统，促进技术、数据、应用和服务的共享，将成为提升竞争力的重要途径。这包括与芯片制造商、硬件供应商、软件开发商、行业应用提供商等多方面的深度合作，形成共生共赢的生态体系。例如端侧大模型推动AI芯片市场发展，2023年全球边缘AI芯片出货预计达22.86亿颗。此外，创新的合作模式，如联合研发、数据共享协议、灵活的IP授权方式等，将促进资源优化配置，加速技术产品的迭代与市场拓展。

来源：企业官网，头豹研究院

中国端侧大模型市场探析——发展趋势

- 未来端侧大模型将在竞争与协作中与云端大模型共同推动AI技术的进步，受手机、自动驾驶和机器人等领域需求的驱动，端侧大模型将不断优化技术，并在个性化、实时响应、隐私保护等方面发挥重要作用

中国端侧大模型行业——发展趋势分析



未来，端侧大模型与云端大模型之间的竞争将进一步加剧

随着计算硬件的持续进步和模型压缩技术的不断发展，端侧大模型在性能、成本、能耗和隐私保护方面的优势将愈发凸显。端侧大模型能够提供更低的延迟和更高的实时响应能力，尤其在网络连接不稳定或高峰期时表现优异。随着用户对数据隐私保护的重视日益增强，端侧大模型的本地处理方式也将成为吸引用户的重要优势。同时，云端大模型在处理复杂计算任务和大规模数据分析方面仍具有不可替代的优势，两者将在不同应用场景中相辅相成，共同推动AI技术的进步。

手机、自动驾驶和机器人等领域的需求将成为端侧大模型发展的重要推动力

在手机领域，个性化服务和实时响应需求不断增加，端侧大模型能够提供智能助手、健康监测和个性化推荐等功能，同时保护用户数据隐私。自动驾驶领域对实时决策和高可靠性的需求将推动端侧大模型的低延迟和本地数据处理能力的进一步提升，从而增强自动驾驶系统的安全性和稳定性。机器人领域的多样化应用场景将促使端侧大模型提供更多自主导航、任务执行和分布式协作的能力，以满足机器人在复杂环境中的智能化需求。

未来，端侧大模型的发展将在技术、应用和生态系统建设等方面呈现多重趋势

在技术方面，模型压缩、硬件加速和边缘计算技术将持续突破，使得端侧大模型在更广泛的设备上能够高效运行。在应用方面，随着智能手机、自动驾驶和机器人等领域对实时响应、低能耗和数据隐私的需求增加，端侧大模型将成为重要的技术支撑。同时，联邦学习、差分隐私等隐私保护技术的应用将进一步提升端侧大模型的安全性和用户信任度。在生态系统建设方面，产业合作和标准化推进将促进端侧大模型的广泛应用和互操作性，推动整个AI产业的健康发展。

来源：企业官网，头豹研究院

Chapter 3
典型厂商分析

- 阿里云
- 商汤科技
- 面壁智能

中国端侧大模型行业典型企业分析——阿里云

- 阿里云在端侧大模型行业的优势主要体现在高效能、低门槛和自适应性三个方面。这些优势使得阿里云能够更好地满足企业和个人对于人工智能技术的需求，推动人工智能技术的发展和应用

阿里云-产品分析

阶段一：模型瘦身

- 量化（权重/激活值）
- 剪枝
- 知识蒸馏
- 混合精度训练
- QAT量化训练
- 稀疏矩阵

阶段二：适配优化

- 高效推理引擎
- 模型分割优化
- 模型芯片适配
- 早期退出
- 级联推理
- 动态网络
- QLoRA

阶段三：产品工程

- 接口设计
- 终端管理
- 模型管理
- 插件管理
- 安全模块
- 端侧RAG
- 升级服务
- 数据融合服务
- 应用使能服务

✓ 竞争优势

01
高效能
阿里云大模型具有更低的成本和更快的开发时间

02
低门槛
阿里云AI大模型提供了丰富的工具库和集成功能

03
自适应性
能够自动调整针对不同任务的选择，提高AI的自适应性

- **阿里云与MediaTek的深度合作是其在端侧大模型领域的重要优势之一**

双方在天玑9300移动平台上完成了通义千问大模型小尺寸版本的端侧部署，并且该部署可以适配天玑8300移动平台，能够实现离线状态下即时且精准的多轮人机对话问答。这种合作不仅推动了AI智能体应用的发展，还为应用开发者和终端设备厂商打造了生成式AI软硬件生态系统。

- **阿里云正在面向AI时代进行全面的技术升级和创新，致力于打造AI时代最开放的云**

通过从底层算力到AI平台再到模型服务的全栈技术创新，阿里云不断提升其在端侧大模型领域的竞争力。此外，阿里云还通过开放的Autonomous Cloud开启AI新时代，进一步巩固其在端侧大模型行业的领先地位。

阿里云在端侧大模型行业的优势主要包括深度合作带来的技术创新、低延时与数据隐私保护以及全面的技术升级和创新，这些因素共同推动了其在该领域的领先地位。

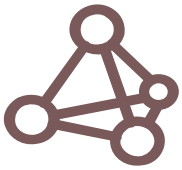
来源：阿里云官网，头豹研究院

中国端侧大模型行业典型企业分析——商汤科技

- 商汤科技在端侧大模型行业的优势主要体现在技术领先与性能卓越、广泛的应用场景与解决方案等方面，使得商汤科技在端侧大模型领域保持领先地位，为各行业提供高效智能的解决方案

商汤科技-产品分析

采用
混合专家框架



基于超过 10TB tokens 训练，大量合成数据

推理上下文窗口
200K

知识、推理、数学、代码
全面对标GPT-4 Turbo

✓ 竞争优势

01

技术领先与性能卓越

语言能力在中端性能手机上实现18.3字/秒的推理速度

02

强大的多模态能力

多模态能力组合赋能产业升级，在多行业都有广泛应用

03

持续的创新与迭代

在知识、数学、推理和代码能力方面都有大幅提升

■ 商汤科技研发的在性能方面表现优异

日日新5.0端侧大模型的语言模型能力在中端性能手机上实现了18.3字/秒的推理速度，而在高端旗舰手机上更是达到了78.3字/秒的推理速度，展现了业内最快的推理速度。在多模态图文能力方面，扩散模型在高通旗舰平台上的推理速度小于1.5秒，比手机云端处理快10倍，且支持输出高清图片和多种图像编辑功能。

■ 广泛的应用场景与解决方案

商汤科技已经通过其强大的多模态能力组合赋能了产业升级，在多个行业都有广泛应用。例如，在金融领域，利用数字人进行智能客服、智慧营销，提供投研分析、研报撰写等新功能，实现降本增效。商汤科技的端侧大模型不仅仅关注单一应用场景，而是通过其多模态能力和高性能，为各种终端任务提供了灵活的解决方案。

■ 模型能力大幅提升

日日新5.0端侧大模型在知识、数学、推理和代码能力方面都有大幅提升，综合能力全面对标GPT-4 Turbo。商汤科技还通过其大模型能力的KRE三层架构（知识-推理-执行）不断突破大模型能力边界，推动AI技术的不断进步。

来源：商汤科技官网，头豹研究院

中国端侧大模型行业典型企业分析——面壁智能

- 面壁智能在端侧大模型行业具有高效推理能力、创新性整合多模态能力、“以小博大”的性能表现、高效低成本、广泛的设备兼容性以及强大的知识库和逻辑处理能力等优势

面壁智能-产品分析

✓ **竞争优势**

01
通用场景能力强
比肩Gemini-Pro、GPT-4V，并超越了全系13B量级模型

02
高效的成本控制
通过全流程高效infra10倍推理加速，将成本降低90%

03
多模态能力
MiniCPM系列模型支持多模态图像视频场景

高效的端侧推理能力和多模态能力的创新性整合

面壁智能发布的MiniCPM端侧大模型支持消费级显卡如英伟达1080Ti的参数微调，以及主流手机处理器的端侧推理。这种能力使得模型能够更高效地运行在终端设备上，提高了用户体验。MiniCPM不仅在端侧加入了多模态能力，而且在实际应用中展现了令人难以置信的能力。从语言翻译到代码编写，再到野外生存指南，其表现均超出预期。这种多模态能力的整合使得模型在处理复杂任务时更加灵活和高效。

虽然MiniCPM的参数规模仅20亿，但其性能表现却十分出色。在多项主流评测榜单上，其中英文平均成绩已超越有着“开源模型新王者”之称的Mistral-7B。这种“以小博大”的特性使得模型在部署和运行时更加节省资源和成本。面壁智能通过技术层面的优化，实现了全流程高效AI infra（推理框架）10倍推理加速，成本可降低90%。这意味着使用MiniCPM进行推理时，能够以更低的成本获得更高的性能。

MiniCPM已跑通了国际主流手机品牌和终端CPU芯片，包括iOS、Android、HarmonyOS等主流操作系统的适配。这使得模型能够更广泛地应用于各种终端设备中。

来源：面壁智能官网，头豹研究院

方法论

- 头豹研究院布局中国市场，深入研究19大行业，持续跟踪532个垂直行业的市场变化，已沉淀超过100万行业研究价值数据元素，完成超过1万个独立的研究咨询项目。

- 研究院依托中国活跃的经济环境，研究内容覆盖整个行业的发展周期，伴随着行业中企业的创立，发展，扩张，到企业走向上市及上市后的成熟期，研究院的各行业研究员探索和评估行业中多变的产业模式，企业的商业模式和运营模式，以专业的视野解读行业的沿革。

- 研究院融合传统与新型的研究方法，采用自主研发的算法，结合行业交叉的大数据，以多元化的调研方法，挖掘定量数据背后的逻辑，分析定性内容背后的观点，客观和真实地阐述行业的现状，前瞻性地预测行业未来的发展趋势，在研究院的每一份研究报告中，完整地呈现行业的过去，现在和未来。

- 研究院密切关注行业发展最新动向，报告内容及数据会随着行业发展、技术革新、竞争格局变化、政策法规颁布、市场调研深入，保持不断更新与优化。

- 研究院秉承匠心研究，砥砺前行的宗旨，从战略的角度分析行业，从执行的层面阅读行业，为每一个行业的报告阅读者提供值得品鉴的研究报告。

法律声明

业务合作

会员账号

可阅读全部原创报告和百万数据，提供PC及移动端，方便触达平台内容

定制报告/词条

行企研究多模态搜索引擎及数据库，募投可研、尽调、IRPR等研究咨询

定制白皮书

对产业及细分行业进行现状梳理和趋势洞察，输出全局观深度研究报告

招股书引用

研究覆盖国民经济19+核心产业，内容可授权引用至上市文件、年报

市场地位确认

对客户竞争优势进行评估和证明，助力企业价值提升及品牌影响力传播

云实习课程

依托完善行业研究体系，帮助学生掌握行业研究能力，丰富简历履历

业务热线

袁先生：15999806788

李先生：13080197867