

2024年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

报告提供的任何内容（包括但不限于数据、文字、图表、图像等）均系头豹研究院独有的高度机密性文件（在报告中另行标明出处者除外）。未经头豹研究院事先书面许可，任何人不得以任何方式擅自复制、再造、传播、出版、引用、改编、汇编本报告内容，若有违反上述约定的行为发生，头豹研究院保留采取法律措施、追究相关人员责任的权利。头豹研究院开展的所有商业活动均使用“头豹研究院”或“头豹”的商号、商标，头豹研究院无任何前述名称之外的其他分支机构，也未授权或聘用其他任何第三方代表头豹研究院开展商业活动。

团队介绍

头豹是国内领先的行企研究原创内容平台和创新的数字化研究服务提供商。头豹在中国已布局3大研究院，拥有近百名资深分析师，头豹科创网(www.leadleo.com)拥有20万+注册用户，6,000+行业赛道覆盖及相关研究报告产出。

头豹打造了一系列产品及解决方案，包括数据库服务、行企研报服务、微估值及微尽调自动化产品、财务顾问服务、PR及IR服务，研究课程，以及分析师培训等。诚挚欢迎各界精英与头豹交流合作，请即通过邮件或来电咨询。

报告作者



袁栩聪
首席分析师
oliver.yuan@Leadleo.com



王利华
行业分析师
lihua.wang@leadleo.com

头豹研究院

咨询/合作

网址：www.leadleo.com

电话：15999806788（袁先生）

电话：18916233114（李先生）

深圳市华润置地大厦E座4105室

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

◆ 中国端侧大模型行业概述	6
• 定义与分类	7
• 发展历程	8
• 驱动力	9
• 市场规模	10
◆ 中国端侧大模型行业产业链分析	11
• 产业链	12
• 模型压缩技术	13
• 成本构成	14
• 厂商类型	15
• 行业场景	16
• 业务场景	17
◆ 中国端侧大模型行业分析	19
• 政策分析	20
• 行业壁垒	21
• 竞争格局	22
• 发展趋势	23
◆ 中国端侧大模型行业典型厂商分析	24
• 阿里云	25
• 商汤科技	26
• 面壁智能	27
◆ 方法论及法律声明	28
◆ 业务合作	29

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 and 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

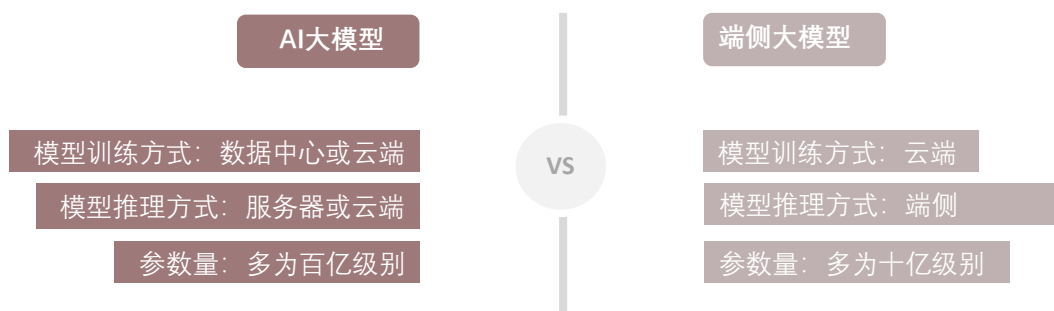
中国端侧大模型市场探析——定义与分类

- 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——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

可靠性、性能和时延

能耗

03

完整版登录 www.leadleo.com

搜索 《2024年中国端侧大模型行业研究》

成本优势

01

端侧大模型

- 从成本优势来看，AI推理的规模越大，单位成本越低。但大型生成式AI模型推理需要大量的计算资源，随着日活用户数量激增，算力需求将呈指数级增长，这将导致规模化部署成本高昂。
- 从能耗来看，端侧大模型推理相比云端推理，功耗更低。端侧设备通常采用低功耗芯片，且推理任务可以在设备本地完成，无需传输数据到云端，从而显著降低能耗。
- 从可靠性来看，端侧大模型推理不受网络波动和服务器故障的影响，能够提供稳定的服务。

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



完整版登录 www.leadleo.com
 搜索 《2024年中国端侧大模型行业研究》

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

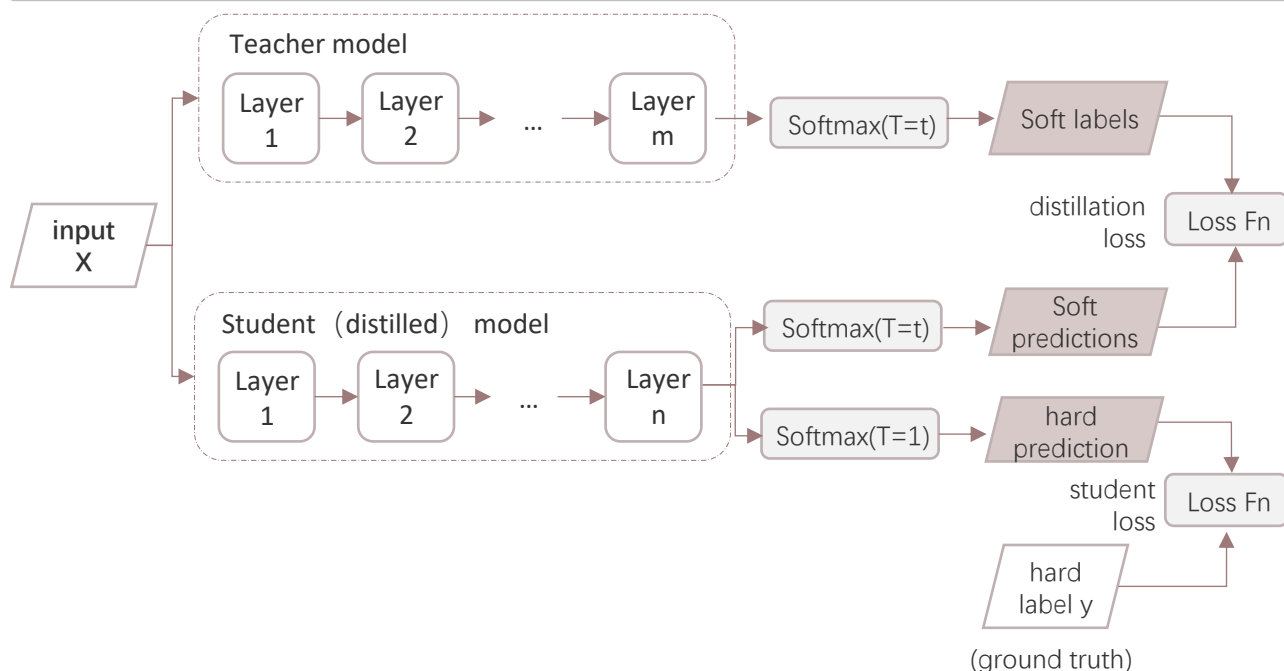


来源：专家访谈，头豹研究院

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



■ 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——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惠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，随着智慧教室的普及率加深，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和设备端，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



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——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

Chapter 3

典型厂商分析

- 阿里云
- 商汤科技
- 面壁智能

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

- ◆ 本报告著作权归头豹所有，未经书面许可，任何机构或个人不得以任何形式翻版、复刻、发表或引用。若征得头豹同意进行引用、刊发的，需在允许的范围内使用，并注明出处为“头豹研究院”，且不得对本报告进行任何有悖原意的引用、删节或修改。
- ◆ 本报告分析师具有专业研究能力，保证报告数据均来自合法合规渠道，观点产出及数据分析基于分析师对行业的客观理解，本报告不受任何第三方授意或影响。
- ◆ 本报告所涉及的观点或信息仅供参考，不构成任何投资建议。本报告仅在相关法律许可的情况下发放，并仅为提供信息而发放，概不构成任何广告。在法律许可的情况下，头豹可能会为报告中提及的企业提供或争取提供投融资或咨询等相关服务。本报告所指的公司或投资标的的价值、价格及投资收入可升可跌。
- ◆ 本报告的部分信息来源于公开资料，头豹对该等信息的准确性、完整性或可靠性不做任何保证。本文所载的资料、意见及推测仅反映头豹于发布本报告当日的判断，过往报告中的描述不应作为日后的表现依据。在不同时期，头豹可发出与本文所载资料、意见及推测不一致的报告和文章。头豹不保证本报告所含信息保持在最新状态。同时，头豹对本报告所含信息可在不发出通知的情形下做出修改，读者应当自行关注相应的更新或修改。任何机构或个人应对其利用本报告的数据、分析、研究、部分或者全部内容所进行的一切活动负责并承担该等活动所导致的任何损失或伤害。

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