

证券研究报告

2023年7月4日

行业报告 | 行业深度研究

数据研究 · 科技专题

AI产业人士看大模型发展趋势

作者：

分析师 孙谦 SAC执业证书编号：S1110521050004

分析师 黄海利 SAC执业证书编号：S1110522090003



行业评级：强于大市（维持评级）
上次评级：强于大市

请务必阅读正文之后的信息披露和免责声明

摘要

人工智能是当今最热门的技术领域之一，也是中国互联网公司的重要战略方向。本报告基于对9位来自中国AI科技团队的产业人士问卷调查，分析了中国AI产业在资源投入、模型发展、数据隐私保护和行业合作等方面的表现，以及面临的挑战和机遇。用科学数据证据给读者提供全面的视角洞察中国AI产业的发展现状和未来趋势。

- **亿级资金有望注入，团队扩容力度加大。**根据公司战略定位和发展重点，在技术研发、算力资源投入、数据采集与标注以及市场推广与商业化扩展方面存在投入差异。同时，AI人力资源也在不断扩张，采取多元化的策略来吸引和培养人才。
- **AI模型新发布可期，复杂数据处理升级。**下半年有多个AI模型发布计划，涵盖自然语言处理、计算机视觉和跨模态领域。在模型发布中，Transformer架构是主流选择。数据挑战、模型优化和商业化仍是AI团队面临的瓶颈。虽然大模型在应用场景中扩展，并非模型规模越大越好，也需综合考虑数据和模型的质量。
- **数据多样性、数据合作和数据隐私保护是中国AI公司在数据领域的关键关注点。**数据多样性与合作是关键，共享数据合作是重要趋势。图像和自然语言数据集普及度高，物体检测数据集应用较少。中国AI公司重视数据安全与隐私保护，采取多层防护措施、动态处理与隐私保护并重，以用户为中心保护用户数据。
- **AI硬件投入将继续保持强劲的发展势头。**服务器部署反映算力需求，大部分公司有服务器扩张计划。不同公司在计算资源的使用量、成本和供应商选择上存在差异，反映出它们在AI技术发展上的投入和战略规划。中国本土公司在半导体领域的发展也不容忽视。
- **AI商业化需要持续投入和优化，而营销策略中突出大模型的创新性和应用价值是至关重要的。**按交易量费和定制开发费是中国AI科技团队主要的收费模式，显示出对需求敏感性和灵活盈利模式的重视。调研结果还揭示了AI服务费用反映了模型复杂性、服务质量和市场竞争的因素，需要综合评估选择。
- **AI的跨行业应用和行业合作是推动技术发展和创新的关键。**AI应用有广阔的发展空间，需要各行业积极与AI公司合作推动数字化和智能化转型，同时加强数据隐私保护。我们认为，未来行业整合、竞争加剧和新兴创业公司崛起的可能性较大。

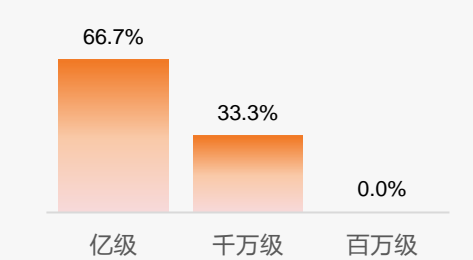
风险提示：样本代表性风险，人工智能行业发展不及预期，商业模式仍不明朗，法律风险

- **亿级资金有望注入与团队扩容力度加大**
 - 2023年亿级资金有望注入，资源投入差异初显布局重点
 - 科技公司AI团队扩容力度加大
- **从架构到发布，从训练到优化**
 - Transformer为主流模型架构，适应复杂任务
 - AI模型新发布可期，复杂数据处理升级
 - 分布式训练与模型并行训练广泛普及，积极探索新的训练技术
 - 模型优化与数据问题是制约模型发展的公认瓶颈
- **数据集的创新、融合和保障**
 - 资料来源多样化、混合化、开放化
 - 图像、语言、问答数据集主导，物体检测集暂露头角
 - 重视数据隐私保护，全方位实践
- **AI发展的底层引擎——计算硬件**
 - 服务器部署反映算力需求，增长意愿仍显热络
 - 计算资源使用量有显著差异，2023年扩增平均幅达20%
 - 2023年AI科技公司计算资源硬件扩增情况
 - GPU单价成本高昂，英伟达为供应商首选，本土公司成长不容小觑
 - 算力战争：硬件和软件相辅相成
- **大模型商业化落地现状与趋势**
 - 按交易量费、定制开发费是主要的收费模式
 - 订阅收费与API收费标准
 - AI科技公司活跃用户总量与月度调用量
 - 大模型垂直应用行业部署与应用成熟度
 - 从“人”“货”“场”看客户拓展策略与成功要素
- **中国AI领域的未来：整合，竞争，开放性与创新**

1 亿级资金有望注入与团队扩容力度加大

2023年亿级资金有望注入，资源投入差异初显布局重点

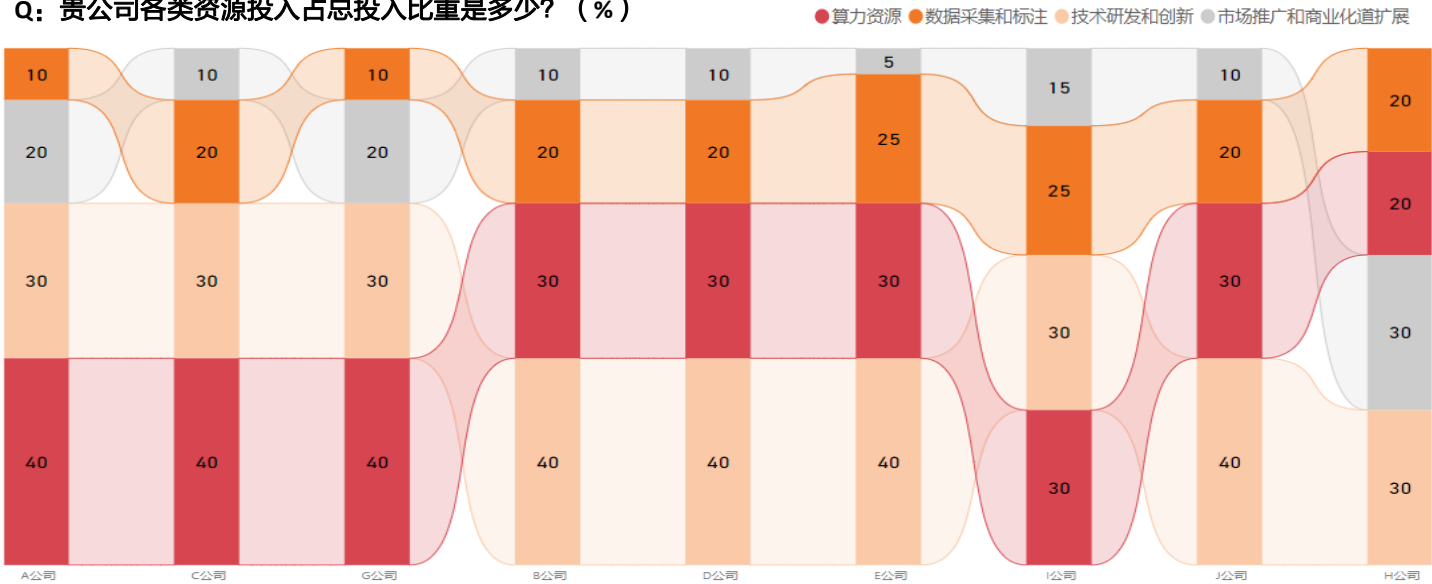
Q：贵公司大模型资金投入是多少量级？



资金投入是AI团队发展的重要保障，66.7%的调研公司在AI领域的投资都达到了亿级规模。

据中国信通院公布的测算数据，2021年中国人工智能产业规模为4041亿元，同比增长33.3%。
据德勤，2020年百度、腾讯、阿里巴巴等企业在人工智能领域的投资金额再创新高，达到1748亿元。

Q：贵公司各类资源投入占总投入比重是多少？（%）

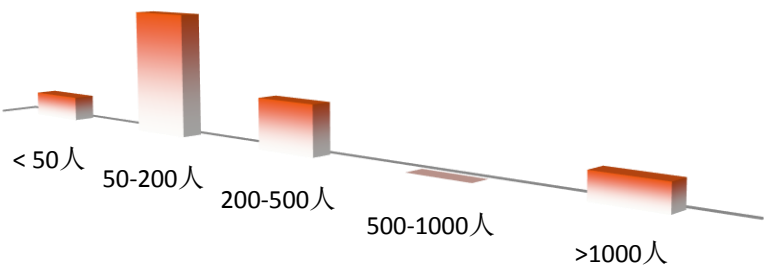


中国AI科技公司在技术研发、算力资源投入、数据采集及标注以及市场推广与商业化扩展方面的投入比重差异，体现了他们的战略定位和发展重点。**技术研发与创新是AI公司持续领先的核心驱动力，占据资源投入的最大比重（avg34%）。**
算力资源投入（avg32%）则是AI研发的基础设施，云计算、分布式计算、AI芯片等在支撑大数据处理和模型训练方面发挥着关键作用。阿里巴巴的阿里云，百度的百度云，华为的云服务等，都在扩充算力资源，以满足AI应用的需求，比如，百度开发了百度机器学习BML(Baidu Machine Learning)平台，提供从开发到部署一站式服务，阿里云为用户提供了阿里云机器学习PAI平台，华为云ModelArts是面向AI开发者的一站式开发平台。
数据采集与标注则是AI算法训练的关键（avg19%）。一些科技公司利用自身的生态系统进行大量的数据采集，并通过人工或半人工方式进行数据标注。例如百度EasyData智能数据服务平台提供便捷的数据采集方案，丰富的数据标注模板及工具，支持将采集、标注、加工等处理后的高质量数据直接对接至EasyDL、BML等百度AI开发平台，服务于后续的模型训练输出更高精度的模型效果。**市场推广与商业化扩展则是AI技术走向市场、实现价值的关键环节。**

科技公司AI团队扩容力度加大

Q：据您了解，贵公司大模型相关工作人员数量？

自人工智能技术开始兴起，中国的科技公司已迅速跻身全球人工智能开发的前列。国内各大科技巨头积极推动AI发展，通过人力资源的扩张和大额资金的注入，以巩固其在市场上的领先地位。人才的重要性不言而喻，是AI创新的主要推动力。



Q：据您了解，贵公司未来是否有人员扩充计划？计划扩容幅度是多少？
采取哪些措施培养AI人才？

	B公司	C公司	A公司	D公司	H公司	G公司	E公司	I公司	J公司
人员扩充计划	√	√	不了解	√	√	√	√	√	√
人员扩张规模	100%+	-	-	10-15%	10%	20%	50%	50%	30%
人才培养措施：									
与高校、研究机构等开展合作培养	√	√	√	√	√		√	√	√
有专门的AI人才培养项目或计划	√		√		√	√			√
在招聘时有针对性的吸引高级AI人才加入		√	√	√			√	√	
提供了具有竞争力的薪酬待遇来吸引或留住高级人才		√	√		√		√	√	
提供外部的专业培训和职业发展机会			√	√		√		√	
设立激励机制来鼓励员工的创新和突破性成果		√	√			√			√
提供内部的专业培训和职业发展机会		√	√						√
关注员工工作生活平衡，为员工创造良好工作环境	√		√						
提供了明确的职业发展路径和晋升机会			√						



88.9%

计划人员扩容

88.9%的参访公司表示未来有人员扩充计划，按照扩充比例分布来看，AI大模型大军扩容激进，其中一家公司近乎人员翻倍的计划。

中国AI科技团队正在采取多元化的策略来吸引和培养人才

- 据与调人士：大部分公司选择与高校或研究机构合作（88.9%），对学生进行实地培训，这种方式既能拓宽人才来源，也可以让人才更早地适应实际工作环境。一部分公司会专门设计一套系统的AI人才培养计划（55.6%），包括提供内外部的专业培训、设立激励机制等，旨在发掘和提升员工的潜力。除了培养内部人才，这些公司在招聘时也会针对高级AI人才制定吸引策略（55.6%），如提供具有竞争力的薪酬待遇。

相对薄弱的培养环节

- 提供内部的专业培训和发展机会（33.3%）：内部培训能够快速提升员工的专业技能，符合公司的发展需求。这表明尚有一些公司在这方面可能有所欠缺。
- 关注员工工作生活平衡，为员工创造良好工作环境（22.2%）：员工的工作满意度和工作效率往往与工作环境和工作生活平衡密切相关。这个比例较低可能意味着许多公司需要更加重视员工的工作生活平衡。
- 提供了明确的职业发展路径和晋升机会（11.1%）：提供明确的职业发展路径和晋升机会能够激发员工的积极性和忠诚度，增强归属感，此比例最低可能反映出一些公司在职业规划方面还有待完善。

2

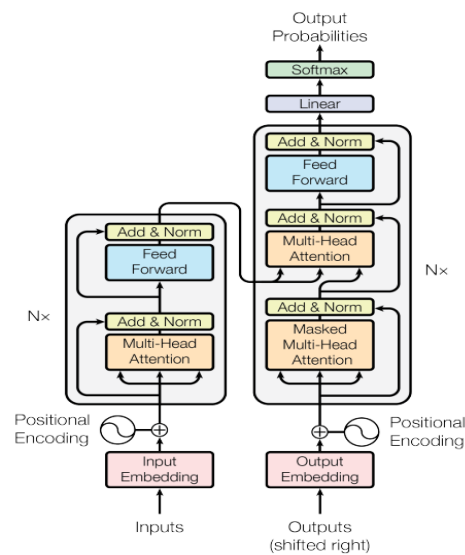
从架构到发布，从训练到优化

Transformer为主流模型架构，适应复杂任务

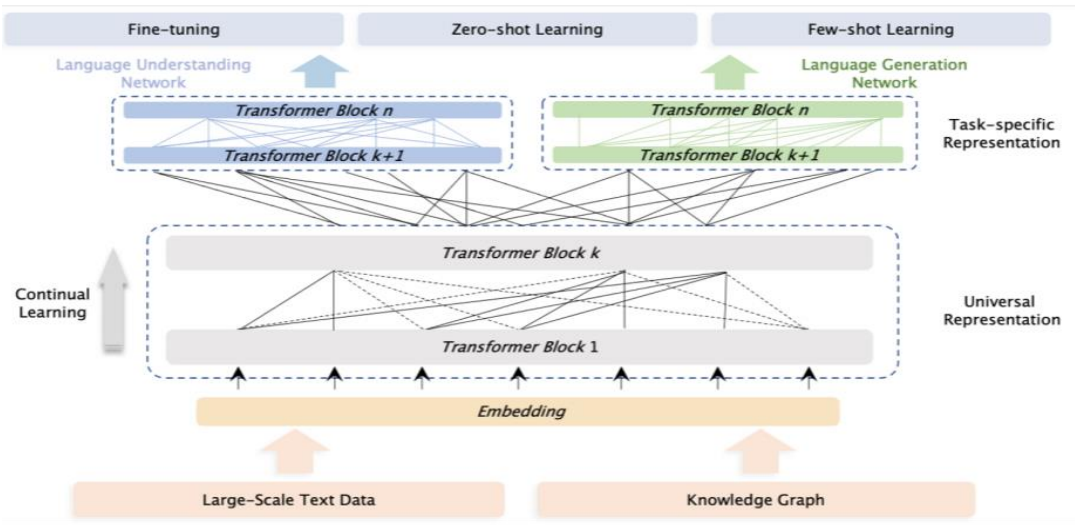
选择合适的模型架构是至关重要的一步，模型的架构决定了它处理数据和学习任务的能力。在对9家中国AI团队的调研中，我们发现**Transformer架构是这些公司最大模型普遍采用的架构**，这一发现揭示了Transformer架构在当下AI领域的重要地位。

在未来的一段时间内，**Transformer可能会保持相对的领导地位**。2017年transformer架构首次被提出，至此以后该架构构成了现代AI训练神经网络的基石，从google的BERT到现在OPEN AI的GPT4，都是基于Transformer的自注意力机制上建立的。纵观国内公司，例如百度在2021年7月5日提出的Ernie便采用了Transformer作为其表示模块，并在该基础上提出了“Continual Multi-Paradigms Unified Pre-training Framework”的预训练框架，并训练出了Ernie3.0,直到2023年3月24日，百度又在基于Ernie和PLATO的基础上训练并推出了NLP大模型文心一言。

Transformer 架构图



百度Ernie3.0架构图



AI模型新发布可期，复杂数据处理升级

- 据与调人士，大模型预计发布的领域主要集中在自然语言处理（NLP）、计算机视觉（CV）和跨模态三个方面。目前，NLP和CV是人工智能领域较成熟和活跃的两个方向，而跨模态是近年来兴起的一个新兴方向。这些领域都拥有丰富的数据资源和多样化的应用场景，为大模型的发展提供了基础和动力。
- 大模型的发布呈现出多样化和细分化的趋势。据与调人士结果，有5家公司计划发布NLP模型（最大参数量级1万亿），6家公司计划发布CV模型（最大参数量级1万亿），4家公司计划发布跨模态模型（最大参数量级1万亿）。有趣的是，调研中的公司都没有科学计算模型发布的计划。
- 大模型的发布频率呈现出加速的趋势。中国的AI团队在2023年开始密集发布各类模型。清华智谱AI研发的GLM-130B 于3月14开启内测，并开源了单卡版模型GLM-6B；百度于3月 16 日推出了其最新的生成式人工智能产品和知识增强型大语言模型（LLM）ERNIE Bot；商汤科技4月10日公布“日日新SenseNova”大模型体系，推出自然语言处理、内容生成、自动化数据标注、自定义模型训练等多种大模型及能力；阿里云4月11日推出语言大模型“通义千问”；科大讯飞5月6日星火认知大模型正式对外发布,时隔一个月，6月9日又推出星火大模型v1.5；北京智源研究院6月9日发布了全面开源的“悟道3.0”系列大模型及算法

Q：据您了解，贵公司2023年是否有发布新模型的计划？预计发布模型参数量是多少？

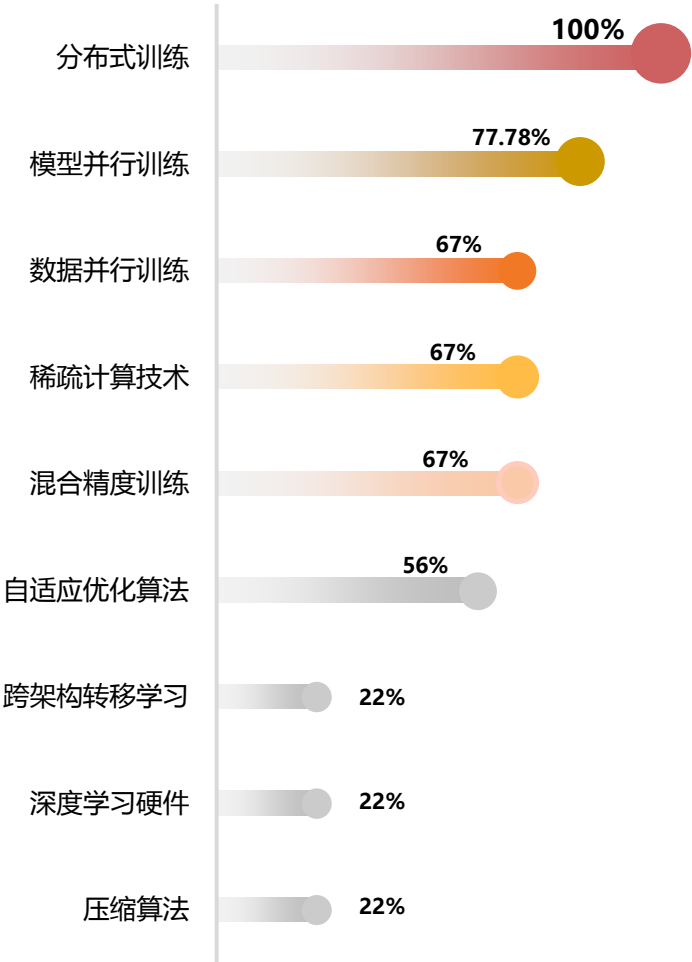
企业	模型发布计划	NLP	CV	跨模态	科学计算
B	√	√	√	√	
C	√		√		
A					
D	√	√			
H	√	√			
G	√		√		
E	√	√	√	√	
I	√		√	√	
J	√	√	√	√	
占比	89%	56%	67%	44%	0%
平均参数量		5.6千亿	2千亿	3千亿	-
最大参数量		1万亿	1万亿	1万亿	-

2023年中国大模型密集发布

发布日期	厂商	模型	类型
3.14	智谱AI	GLM-130B	NLP
3.16	百度	文心一言	NLP
4.1	商汤	日日新	跨模态
4.11	阿里	通义千问	NLP
5.6	科大讯飞	星火认知大模型v1.0	跨模态
6.9	科大讯飞	星火认知大模型v1.5	跨模态
6.9	智源研究院	悟道3.0	NLP

分布式训练与模型并行训练广泛普及，积极探索新的训练技术

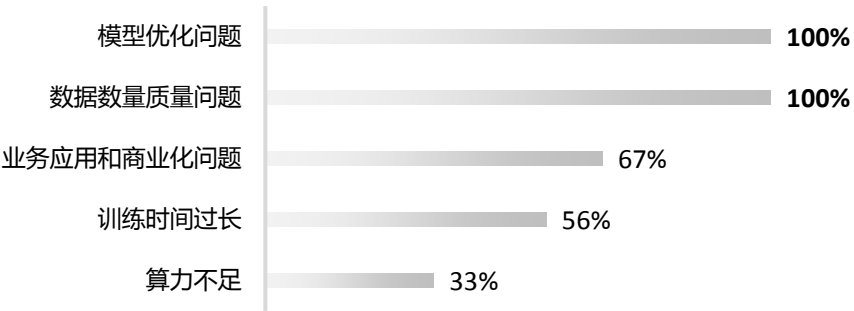
Q：据您了解，贵公司在大型模型训练方面有哪些创新技术或方法？



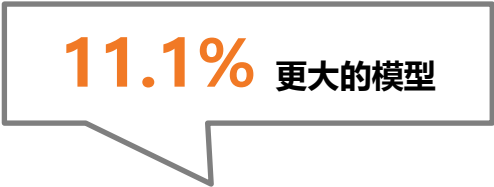
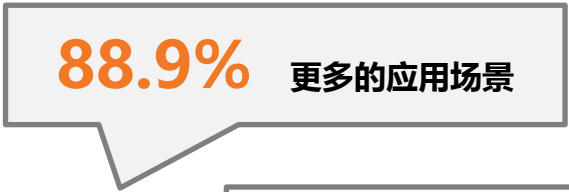
- 分布式训练**
利用数据并行或者模型并行的策略将一个大型的模型划分为多个部分，分配到不同的设备上进行训练，从而提高训练效率和模型规模。**优势：**可以训练超大规模的模型，突破单个设备的内存和计算能力的限制，同时也可以加速训练过程。
- 模型并行训练**
利用图切分算法，指将一个模型的计算图按照层次或者功能划分为多个部分，然后根据子图之间的依赖关系，在不同的设备上按照顺序或者并发地执行。**优势：**处理非常深或者宽的网络结构，克服单个设备内存不足的问题。
- 数据并行训练**
利用数据切分算法，将一个大批量的数据切分为多个小批量的数据，然后根据数据之间的独立性，在不同的设备上并行地执行前向和反向传播。**优势：**提高设备的利用率，加速训练过程，同时也可以增加批量大小，提高模型的泛化能力。
- 稀疏计算技术**
在训练或者推理时利用模型中存在的稀疏性（如零值或者低值）来减少计算量和内存占用的一种技术。**优势：**稀疏计算技术的优势是可以大幅降低模型的存储和计算成本，提高模型的性能和效率。
- 混合精度训练**
在训练时在模型中同时使用16位和32位浮点类型，从而加快运行速度，减少内存使用的一种训练方法。**优势：**利用现代硬件（如NVIDIA GPU或者Cloud TPU）对低精度运算的支持，提高运算速度和吞吐量，同时也可以节省内存空间和带宽，从而增大批量大小或者模型规模。

模型优化与数据问题是制约模型发展的公认瓶颈

Q: 据您了解，贵公司在训练大模型时，主要面临哪些挑战？



Q: 您认为大模型发展最确定的趋势是什么？



数据挑战、模型优化与商业化挑战

- **数据是训练深度学习模型的关键。**在训练深度学习模型时，大量高质量的数据是非常关键的。获取这样的数据通常需要投入大量的时间和资源。此外，数据的清洗和标注也是一个重大的挑战，需要大量的人力进行操作。
- **模型优化是AI团队面临的问题。**AI模型优化主要是为了使模型在不同的硬件平台上快速运行，包括算法层面的优化、框架层面的优化以及硬件层面的优化等手段。
- **商业化受到关注。**业务应用和商业化也逐渐成为科技公司关注的焦点，据与调人士结果，6家公司都面临着同样的问题。

应用拓展重于模型规模

- **大模型的应用场景将持续扩展。**据调研，88.89%的受访人士认为大模型将会有更多的应用场景。这一趋势的背后是深度学习技术在语音识别、图像识别、自然语言处理等领域的突破。我们推断垂直商业化将成为科技公司的主要关注点，仅有11%的公司认为更大规模的模型会成为趋势。
- **参数量不是衡量模型好坏的唯一标准。**模型参数越大，训练的成本也越高。并且在大模型训练领域中，存在着边际效应递减的现象。以谷歌发布的拥有1.6万亿参数的Switch Transformer为例，当谷歌把参数量提升了一个量级后，确实会对性能有所提升，但是此时带来的性能收益已经远不及以前。有时候在更多数据上训练的较小模型表现更好，而不是在较少数据上训练的较大模型。例如，DeepMind的Chinchilla模型拥有700亿个参数，并在1.4万亿个token上进行了训练，而2800亿参数的Gopher模型在3000亿个token上进行了训练。在随后的评估中，Chinchilla的表现优于Gopher。可见参数量级并不能完全决定模型的高低。

3

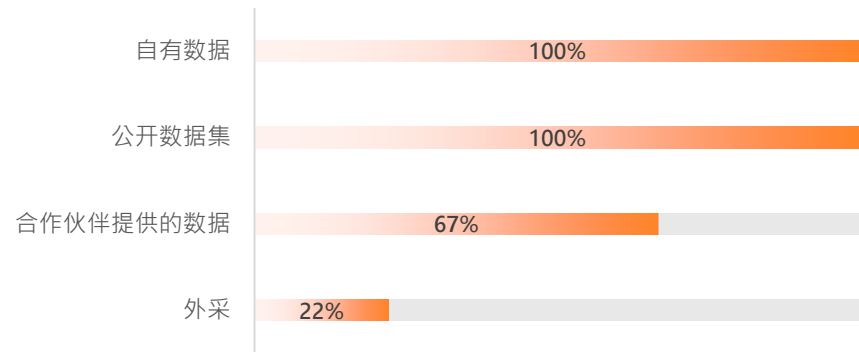
数据集的创新、融合和保障

资料来源多样化、混合化、开放化

公开数据集和自有数据集的使用仍然占据主导，合作伙伴的数据源和外采数据的利用则更侧重于补充和优化。

- **数据多样性与合作是关键。**据与调人士结果，公司都同时采用自有数据集和公开数据集。这意味着他们将自己收集和整理的数据与来自公开资源的数据结合使用，以期在模型训练和优化中获得更好的效果。这种混合使用数据的方式可以增强数据的多样性和全面性，从而提高模型的鲁棒性和泛化能力。
- **共享数据合作是重要趋势。**据与调人士结果，6家公司选择采用合作伙伴提供的数据源（占67%）。公司间数据的交换和合作在AI领域的重要性不容小觑。通过与合作伙伴共享数据，可以扩大数据规模，增加数据的维度和深度，以支持更复杂、更精细的模型训练。例如，Snowflake 数据云平台在全球拥有超过 8,000 家客户（截至2023年4月30日），Snowflake 与 NVIDIA 合作将通过数据云平台把定制化的生成式 AI 应用带到不同的垂直领域，从而进一步帮助客户改变这些行业。
- **外采数据潜力有待挖掘。**据与调人士结果，只有2家公司选择了外采数据，占比为22%。这可能反映了大部分公司依赖于内部和合作伙伴的数据源，或出于数据安全性、质量和可用性等因素的考虑。而较少去寻求外部数据市场的支持。外采可以补充公司自身难以获得的数据，从而优化模型效果。
- D公司在数据获取方面具有较强的能力和意识，资料来源类型最为丰富，同时拥有4类数据源。

Q：据您了解，贵公司目前使用的主要资料来源是什么？



自有数据集：来源于公司自身的产品和服务；这些数据集通常比公开数据集更加具有针对性和实用性。混合数据集则结合了公开数据集和自有数据集的优点，可以在保持模型泛化性能的同时，满足特定任务的需求。

公开数据集：从公开渠道获取的数据集；具有较大的规模和多样性，但质量相对较低，可能存在噪声、偏差或过时等问题。

合作伙伴提供的数据源：与其他机构或企业进行合作或交换而获得的数据源；通常是针对相关领域或任务的数据，具有较高的可靠性和实用性，但获取条件较为苛刻，可能存在隐私、安全或法律等风险。

外采数据：从第三方机构或平台购买或租赁而获得的数据；巨有较高的补充性通常是针对缺失或不足的。

图像、语言、问答数据集主导，物体检测集暂露头角

Q：据您了解，贵公司使用的数据集类型是什么？

			
图像数据集 图像或视频	自然语言数据集 文本或语音	问答数据集 问题和答案	物体检测数据集 物体位置类别
计算机视觉领域或任务，如图像分类、目标检测、人脸识别、视频理解等。	自然语言处理领域或任务的数据，如文本分类、情感分析、机器翻译、语音识别等。	问答系统领域或任务的数据，如知识问答、阅读理解、对话系统等。	物体检测领域或任务的数据，如行人检测、车辆检测、行为识别等。
具有较高的多样性和规模，但也需要较高的处理和标注成本。	具有较高的复杂性和丰富性，但也需要较高的理解和分析能力	具有较高的实用性和价值，但也需要较高的逻辑和推理能力。	具有较高的精确性和难度，但也需要较高的计算和存储资源。
<ul style="list-style-type: none">LabelmeImageNetLSUNMS COCO	<ul style="list-style-type: none">Enron DatasetAmazon ReviewGoogle Books Ngrams	<ul style="list-style-type: none">cMedQA 2.0Chinese-Medical-Question-Answering-System	<ul style="list-style-type: none">Objects365VEDAIDOTA

数据集决定模型性能与适用范围

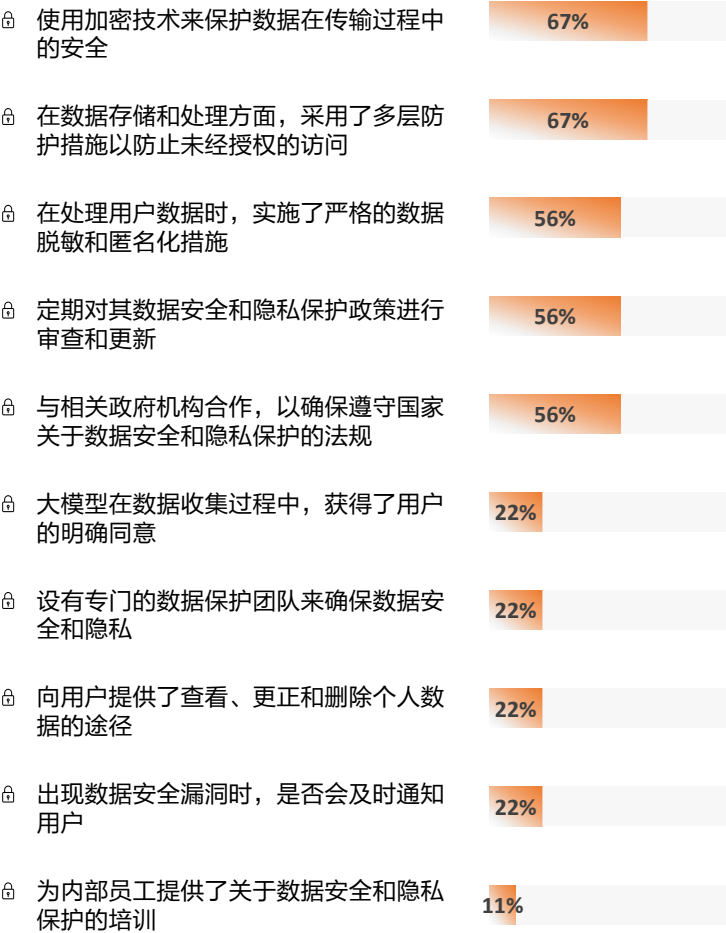
- 图像和自然语言数据集普及度高。**图像数据集、自然语言数据集和问答数据集在这些公司中使用较为普及，占比达到了88.9%。我们认为原因可能是，视觉识别和自然语言处理是主要关注的AI研究方向。同时，这也反映了图像和自然语言数据在AI应用中的重要性。
- 物体检测数据集应用较少。**据与调人士结果，仅5家公司使用物体检测数据集。物体检测在自动驾驶、安全监控等领域有重要应用，这表明上述领域可能并非所有公司的重点研究方向，或者他们正在寻找其他类型的数据来训练相关模型。
- 不同企业数据集规模有所差异。**从数据集的规模来看，各企业数据集优势有所差异。据与调人士结果，B企业重图像和自然语言数据集，C企业重图像和问答数据集，A企业问答和物体检测数据集较为丰富，D企业在图像和问答数据集较有优势。这体现了不同公司依据其业务方向和专业优势选择数据集类型，以满足特定应用需求。

Q：贵公司使用的数据集规模有多大？

企业	图像数据集 万张图像	自然语言数据集 TB文本	问答数据集 万个问答	物体检测数据集 万个物体类别
B	40000	500+		
C	15000-20000	1.5	7000-8000	
A	千万级	5	亿级	亿级
D	100000	20	2000000	2
H	50	1	100	10
G	100	0.5	10	20
E	500	200	50	25
I			50	
J	80	5	20	

重视数据隐私保护，全方位实践

Q：据您了解，公司在数据安全、隐私保护等方面做了以下哪些措施？



中国AI公司在数据安全与隐私保护方面表现出高度的重视与严谨的态度，采取了一系列有效的措施：

- **多层防护措施保障数据安全性。**据与调人士，大多数公司采取了多层防护措施来防止数据在存储与处理过程中的未经授权的访问，比如使用加密技术来保护数据在传输过程中的安全。显示了中国AI公司对于数据安全性的高度重视，以及在技术手段上的实力与应对能力。
- **动态处理与隐私保护并重。**据与调人士，半数的公司定期对数据安全与隐私保护政策进行审查与更新，并且在处理用户数据时，实施了严格的数据脱敏与匿名化措施。这种动态的、严格的处理方式，既保证了数据的使用效率，又兼顾了用户隐私的保护。
- **用户为中心，保护用户数据。**据与调人士，22%的公司在数据收集过程中，得到了用户的明确同意，并设立了专门的数据保护团队来确保数据安全与隐私。用户在任何时候都可以查看、更正和删除个人数据，同时，一旦出现数据安全漏洞，这些公司也会及时通知用户。这种以用户为中心的方式，不仅符合了法律规定，也提高了用户对公司的信任度。
- **员工培训促进安全文化。**据与调人士，11%的公司为内部员工提供了关于数据安全与隐私保护的培训，这一举措显示了这些公司对于创建安全文化的重视，也为员工提供了必要的知识与技能，以更好地保护用户数据的安全与隐私。

国家法规介入，AIGC数据安全需重视

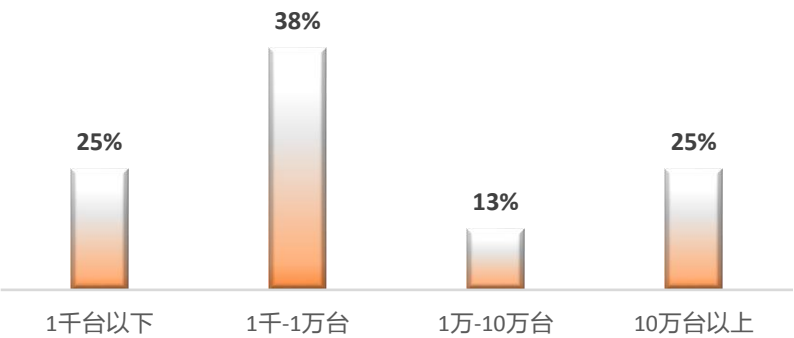
- 在大模型的训练中应当通过数据安全技术保证所取用户的个人隐私信息不被其他终端获取或应用于其他用途。2023年5月10日国家网信办起草了《生成式人工智能服务管理办法（征求意见稿）》，其中第五条明确表示了利用生成式人工智能产品提供聊天和文本、图像、声音生成等服务承担该产品生成内容生产者的责任，另外涉及到个人信息的，也应当承担个人信息处理者的法定责任，履行个人信息保护义务。这在法规上要求了未来AIGC的内容应当具有稳定性和准确性，并且保证用户信息的隐私性。

4

AI发展的底层引擎——计算硬件

服务器部署反映算力需求，增长意愿仍显热络

Q：据您了解，贵公司大模型服务器使用数量是多少？今年是否有增加服务器的计划？



企业	服务器数量（台）	今年计划扩增数量（台）
B	100000	100000
C	160000	2000
A	20000	30000
G	2000	1000
I	2000	10000
D	1000-2000	1000
E	1000-2000	2000
H	500	1000
J	200-300	-

人工智能技术的发展在很大程度上取决于算力的支持，而服务器是提供算力的关键设施。本调研结果显示，中国AI科技团队都在大规模地部署和使用服务器，且大部分公司仍有服务器扩张的计划，并且扩张幅度都不小。据与调人士：

服务器拥有情况：

- 大多数公司的AI服务器数量在1万台以下。
- B公司和C公司拥有服务器数量位居头部，远超其他公司。
- 其他公司主要分布在2000台以内。

服务器的扩张计划：

- B公司、A公司和I公司的扩张计划最为激进，计划增加万台以上的服务器数量。
- 其他公司计划扩张1000-2000台的规模。

服务器数量的多少和扩张计划的存在，都反映出这些公司在AI技术发展上的投入和决心。特别是互联网巨头，他们的服务器数量和扩张计划显示出其在AI领域的领先地位和雄心壮志。而其他公司，尽管服务器数量相对较少，但也通过扩张计划展示了对AI发展的期待和支持。我们认为，调研结果揭示了中国AI科技公司在算力部署上的实力和意愿，预示了中国AI技术将有望继续保持强劲的发展势头。

计算资源使用量有显著差异，2023年扩增平均幅预达25%

Q：据您了解，贵公司最大模型用到以下哪些计算资源硬件？

GPU 100%

(图形处理单元)在AI领域具有重要的作用，特别是在处理深度学习和机器学习任务时。GPU可以并行处理大量的计算任务，这使得它们在处理图像和视频数据时具有显著的优势。

CPU 100%

(中央处理单元)在处理更为复杂和多变的计算任务时，CPU具有更好的灵活性。在AI领域，CPU常常被用于处理不适合在GPU或TPU上运行的任务，或是在资源有限的环境下运行轻量级的模型。

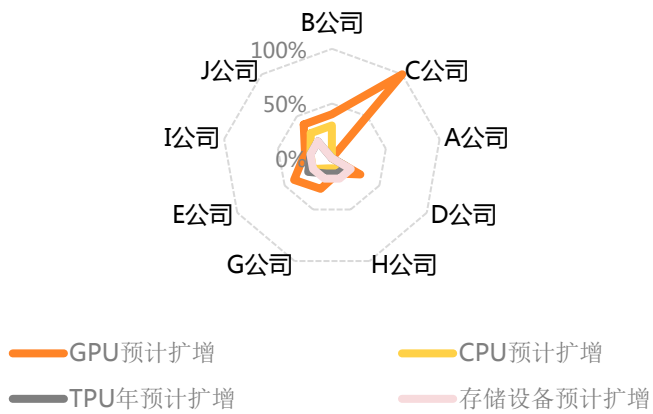
TPU 78%

(张量处理单元)是Google专门为机器学习工作负载设计的处理器。TPU专为大规模矩阵运算和高吞吐量的低延迟运算优化。

存储设备 89%

存储设备可以缓存AI大模型的参数和中间结果，减少内存访问损失和计算开销，提高模型的推理速度。

Q：据您了解，2023年贵公司预计增加计算资源硬件的幅度是多少？



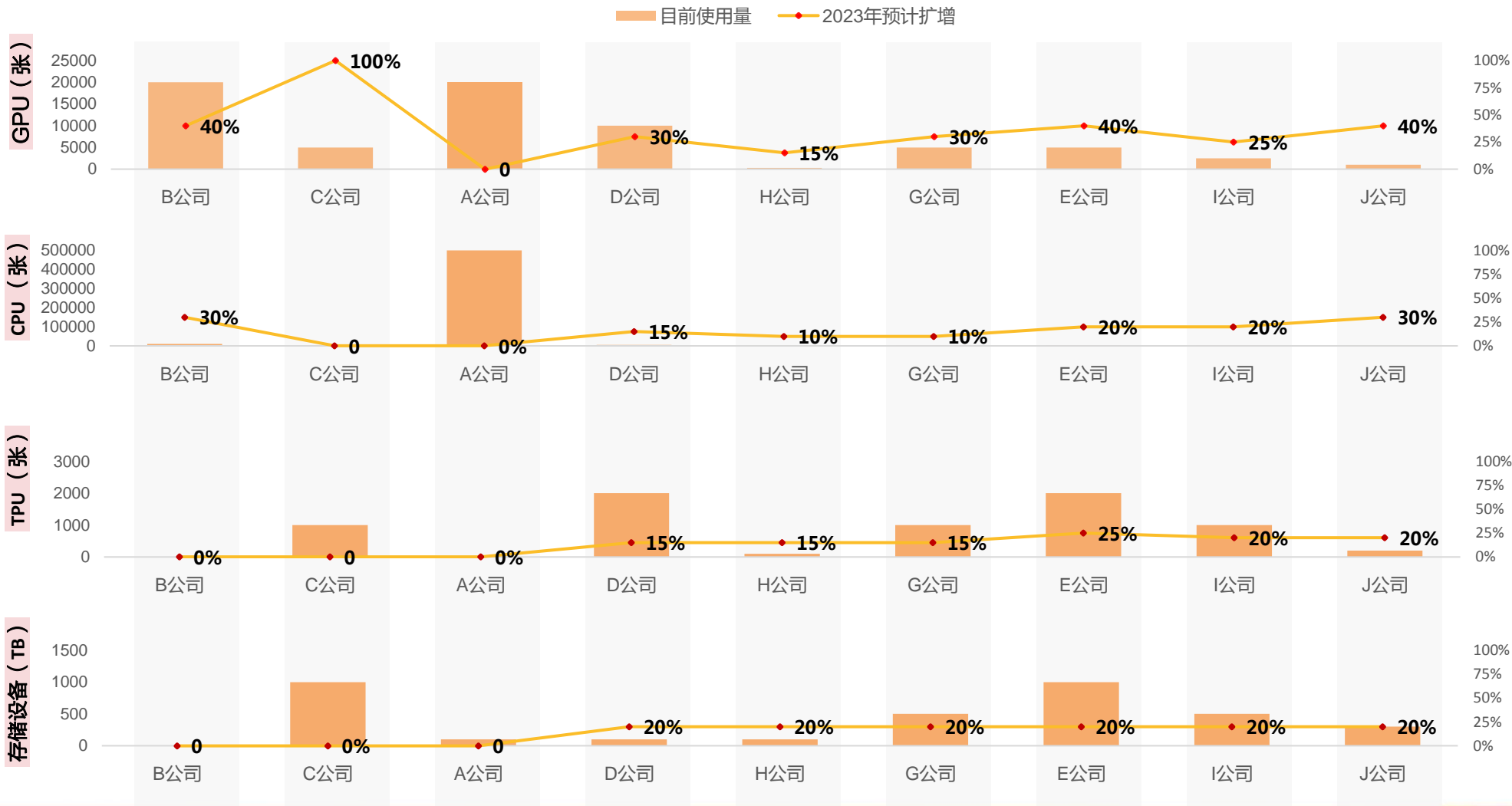
计算资源直接影响到模型的训练效率和实用性，重要性不言而喻。

- **GPU：** 仍然是最主流的计算硬件之一。尤其是对于互联网和云服务等领域企业，GPU可以提供高性能和高效率的AI训练和推理服务。据与调人士，B公司、A公司、D公司在这方面表现突出，其GPU使用数量均超过1万张，明显高于其他公司。而在未来的扩增计划中，8家公司表示将有所增加（平均幅度40%），其中C公司的扩增计划最为激进，预计GPU数量将翻倍，这表明他们对AI技术的投资意愿强烈。
- **CPU：** 使用情况分化明显。据与调人士，A公司和B公司同样表现出强大的实力，其CPU数量均达到万级。在未来的扩增计划中，平均增幅为19%。B公司和J公司的计划较为激进，预计扩增30%，显示出这两家公司在大数据处理和高性能计算领域的野心。
- **TPU：** 使用相对较少。据与调人士，D公司和E公司的数量约是平均值的两倍，这两家公司在机器学习和深度学习方面的需求可能是导致TPU使用较多的主要原因。在未来扩增计划上，E公司预计会增加25%，远高于其他公司，反映了其在深度学习领域的发展计划。
- **存储设备：** 使用分布较为均衡。尤其是对于数据密集型的AI应用场景，存储设备可以提供快速和安全的数据访问和传输。据与调人士，C公司和E公司的存储设备数量约是平均值的两倍，显示了他们在大数据处理和存储方面的强大能力。在未来的扩增计划中，多数公司都表示将有所增加，平均增长率约为20%，我们认为这可能表明，他们都认识到了大数据时代数据存储设备的重要性。

随着AI技术的发展，我们预期这些公司会继续扩大对这些计算资源的投入，并根据自身的业务需求和战略方向进行调整。

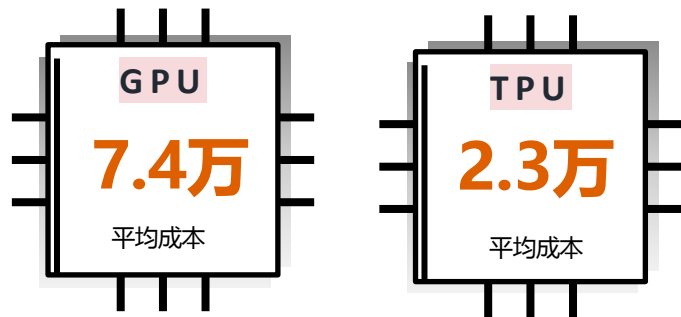
2023年AI科技公司计算资源硬件扩增情况

Q: 据您了解，2023年贵公司预计会增加以下哪些计算资源？



GPU单价成本高昂，英伟达为供应商首选，本土公司成长不容小觑







Q：据您了解，贵公司GPU和TPU的单价成本是多少元？



高效处理并行运算，GPU投入不容小觑：据与调人士，企业平均的GPU和TPU成本分别为7.39万元和2.29万元，尽管GPU的成本较高，但其在处理并行运算，尤其是深度学习算法方面的性能表现卓越，使得这一额外的投入成为企业无法避免的支出。

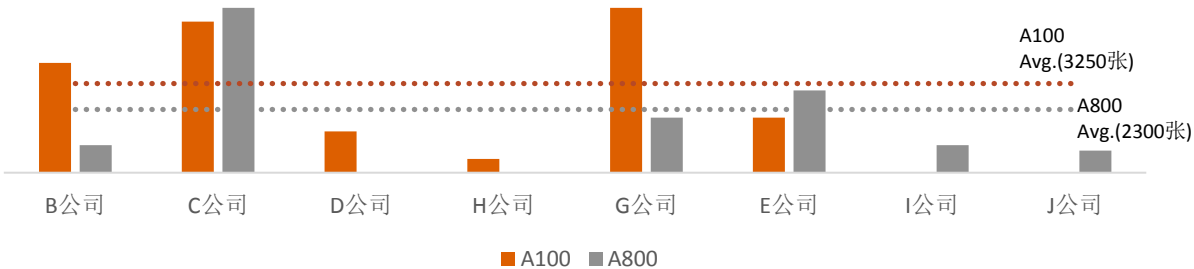
Nvidia GPU主导市场，多元选择依然存在：从市场占有率来看，GPU仍然是深度学习中最受欢迎的处理器架构。Nvidia在GPU领域具有较强的竞争优势和品牌影响力，实际应用中多元化的供应商选择依然存在。据与调人士，所有9家公司都选择了nvidia的GPU作为主要方案，AMD的GPU也得到了一些公司（C、D、H）的青睐。

Q：据您了解，贵公司采用的GPU是由哪家供应商提供？

	100 %
	33.3 %
	33.3 %
	22.2 %
	11.1 %
	11.1 %

华为和寒武纪，国内GPU市场崭露头角：中国国内的华为和寒武纪也开始在GPU市场中崭露头角。他们的产品分别被两家公司选择为GPU供应商。

NVIDIA A100和H100受青睐，C公司投入领先：在具体型号上，NVIDIA的A100和H100平均持有量分别是3250张和2300张，这也证明了这两款产品在市场上的普遍认可。特别是C公司表现出在硬件投入方面的雄心壮志，他们目前拥有的A100和H100总量最高（总量超过1万张）。



算力战争：硬件和软件相辅相成

国内外各型号AI芯片算力表

	供应商	型号	INT8	FP16
国内	华为	昇腾910	640TOPS	320TFLOPS
	寒武纪	思元290	512 TOPS	-
	百度	昆仑芯二代AI芯片	-	128 TFLOPS
国外	Nvidia	A100 80GB PCIe	624 TOPS	312 TFLOPS
		A100 80GB SXM	1248 TOPS	624 TFLOPS
		H100 SXM	3,958 TOPS	1,979 TFLOPS
		H100 PCIe	3,026 TOPS	1,513 TFLOPS
		H100 NVL	7,916 TOPS	3,958 TFLOPS
	AMD	Instinct MI250	362.1 TOPs	362.1 TFLOPs
		Instinct MI250X	383 TOPs	383 TFLOPs
		Instinct MI300	计划今年下半年推出，预计有MI250X系列8倍的人工智能训练性能	

华为昇腾计算能力追上A100 PCIe1 CUDA软件生态打造Nvidia护城河

- 在国内AI芯片市场，单卡AI芯片算力最高的是华为旗下海思的昇腾910，在半精度下可以达到320TFLOPS的计算速度，与Nvidia的A100 PCIe版本持平。
- 在国外AI芯片市场，Nvidia遥遥领先，其H100 NVL版本在半精度下算力可以达到3958TFLOPS，是AMD Instinct MI250X的10.33倍，华为昇腾910的12.4倍，在算力方面Nvidia在全球市场的地位显而易见。
- 根据AMD官网发布消息，今年下半年预计推出的Instinct MI300有8倍于MI250X的计算性能，旨在挑战Nvidia的H100系列。
- 另外，虽然AMD的MI300系列在算力速度上跟上了Nvidia系列，但是上层的软件架构同样重要，过去十年取得的AI进步大部分是通过CUDA库完成的，其他厂商想要挑战Nvidia的主导地位除了算力方向的提升，同样要加强软件生态的建设，这是一个投入巨大并且漫长的过程，在短时间内Nvidia市场地位无法撼动。

5

大模型商业化落地现状与趋势

按交易量费、定制开发费是主要的收费模式

Q: 据您了解，贵公司大模型商业化主要采用哪种商业模式？



交易量收费

➤ 交易量收费

根据客户每月使用的API调用或交易量收取费用。定价标准通常是按交易量计算，例如每千个API调用收取一定的费用。



定制开发收费

➤ 定制开发费用

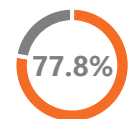
如果客户需要特定领域的AI模型，公司通常会收取定制开发费用。定价标准通常取决于开发的难度和时间成本。



服务收费

➤ 服务费用

提供数据处理、标注和质量控制服务等。



订阅服务收费

➤ 订阅收费

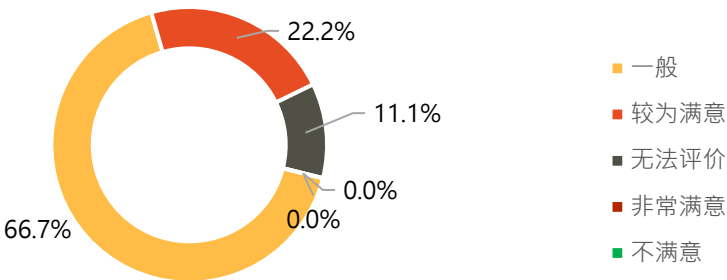
客户可以根据需要选择不同的订阅级别，如基本、标准或高级。订阅费用通常按月或按年收取，并根据所需服务的数量和类型进行定价。

全面覆盖常见的付费模式，凸显需求敏感性和盈利模式灵活性。

- 定制化解决方案是主要收入来源。据与调人士这种收费模式已全面覆盖所有公司，例如美国的OpenAI就采取了类似的商业模式。说明相关AI科技团队主要依靠为客户提供定制化的AI解决方案来获取收入，而不是提供标准化的AI产品或平台。
- 数据相关服务是重要的收入来源。AI科技团队提供AI解决方案的同时，也需要提供数据相关的服务，如数据采集、清洗、标注、质量控制等，这些服务往往需要大量的人力和时间成本。
- 订阅收费模式已兴起。提供基于云端或SaaS的AI服务，体现了这些公司对长期客户的关注，通过提供持续的服务，建立起稳定的客户关系。让客户可以按需使用AI能力，而不是一次性购买。

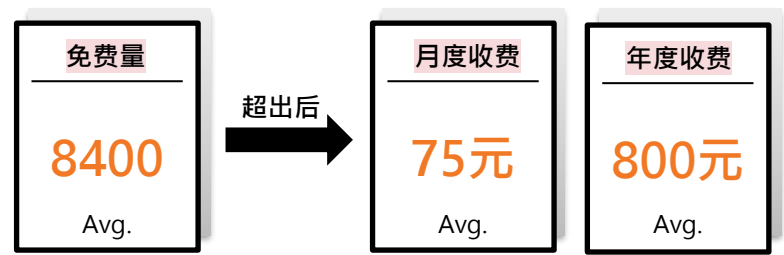
尽管对目前商业化投资回报的评价多为中性，但这并没有阻止他们在商业化道路上的探索。AI商业化是一个长期过程，需要持续投入和优化。

Q: 您认为贵公司大模型商业化投资回报如何？

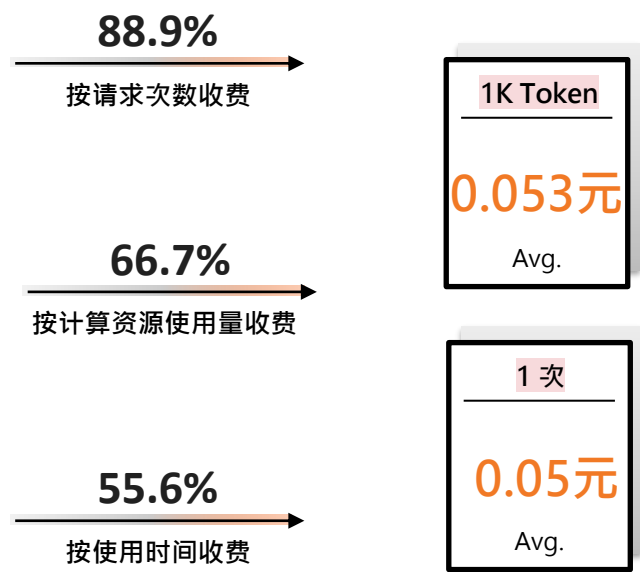


订阅收费与API收费标准

Q: 据您了解，贵公司大模型订阅收费标准是多少？



Q: 据您了解，贵公司交易量收费标准是多少？



AI服务费用反映其模型复杂性、服务质量和市场竞争，需要综合评估选择

大模型在开发和运行所消耗的资源，无论是在算力还是存储空间方面，都是非常庞大的。因此，AI团队需要以某种形式来收回这些资源的投入，从而形成了各种不同的收费标准。对于不同的收费标准进行比对，有助于我们更好的理解和评估各家公司的服务价值。调研结果显示，大多数公司采用的是订阅收费和API收费的模式。

¥ **订阅收费模式：**据与调人士结果，7家团队采用了订阅收费的标准，其中免费量1000到20000次（平均8400次），月度收费平均价格50-100元（平均75元），年度收费价格600-1000元（平均800元）。订阅收费模式能让用户更清楚的知道他们需要支付的费用，并且在一定范围内，可以无限制的使用AI服务，有利于培养用户的付费习惯。

¥ **API收费模式：**据与调人士结果，与调团队都采用了API收费的模式，其中按照请求次数收费的团队占比88.9%，按照计算资源使用量收费占比66.7%，按照使用时间收费占比55.6%。API收费模式为用户提供了灵活性，他们可以根据实际需求来付费。

全球范围内的类似收费模式：Google Cloud、Amazon Web Services（AWS）和Microsoft Azure等。他们的费用通常基于API调用次数、数据处理量或使用的计算资源。具体的费用标准可以在他们的官方网站上找到，但总的来说，他们的费用往往更高，主要是由于他们所使用的模型更加复杂，以及他们的服务通常包括了更多的增值服务。

AI服务费用反映背后因素：无论是在中国还是在全球范围内，AI服务的费用标准通常反映了其背后的模型复杂性、服务质量以及市场竞争程度。因此，选择哪一种服务不仅要看价格，还要考虑其背后的技术支持、服务质量以及具体的需求。

AI科技公司活跃用户总量与月度调用量

Q：据您了解，贵公司API月度用量是多少？

接入量级	百分比	公司	平均月度用量
万亿级	11.1%	C公司	40万亿
亿级	22.2%	B公司、A公司	5.5亿
万级	44.4%	H公司、G公司、I公司、J公司	22.5万
不确定	22.2%	D公司、E公司	-

Q：据您了解，贵公司AI应用活跃用户总量是多少？

活跃用户	百分比	公司	平均活跃用户
千万级	11.1%	G公司	2千万
百万级	22.2%	B公司、A公司	250万
十万级	22.2%	C公司、D公司	50万
万级	33.3%	H公司、J公司、I公司	3万
不确定	11.1%	E公司	-

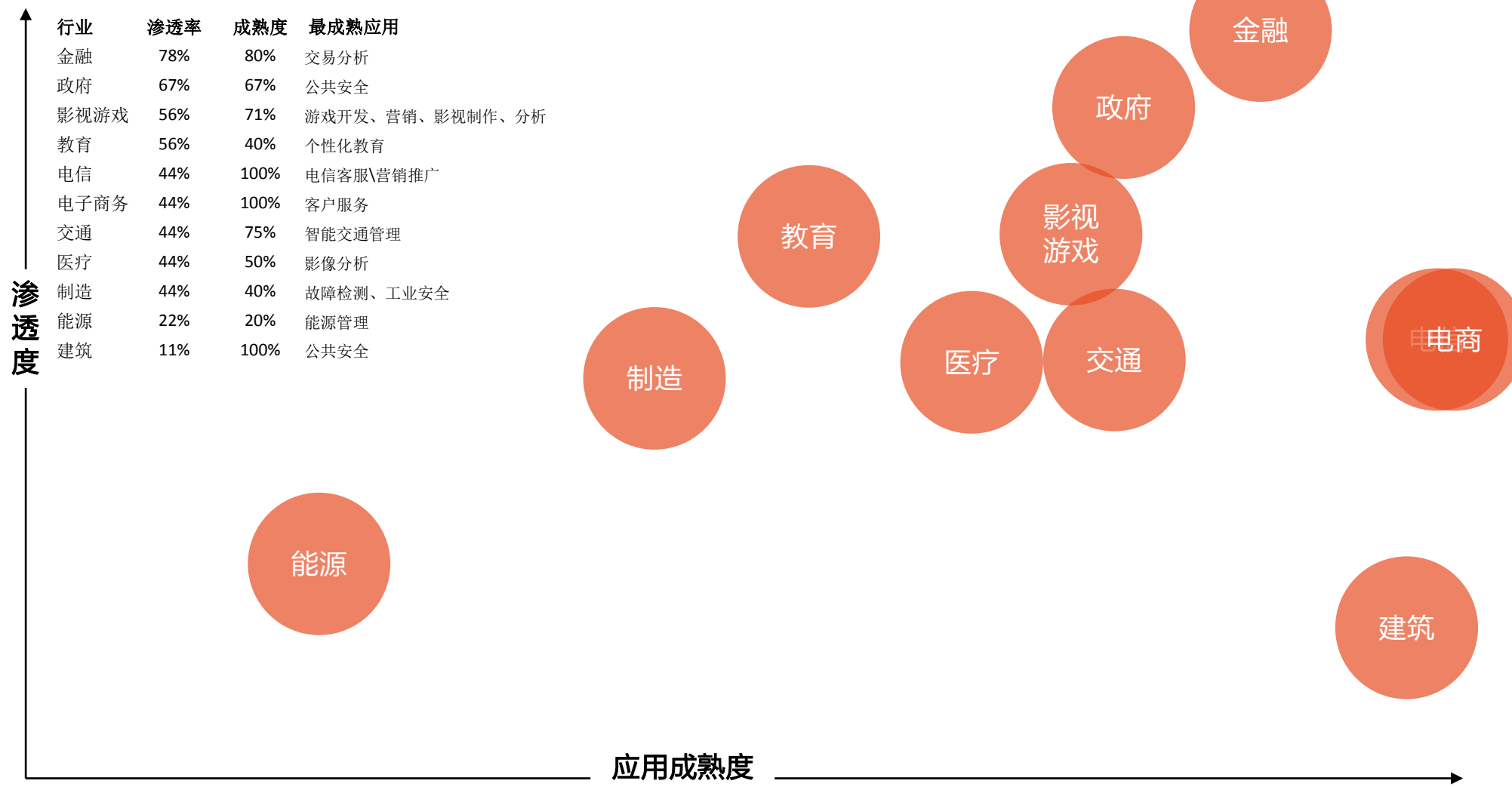
API调用量和用户活跃度的关系

- 针对用户的订阅和企业的API收费大模型商业化的主要收入来源，因此API月度调用量和AI应用活跃用户数是衡量API使用情况和AI应用影响力的重要指标。虽然大规模API调用量通常与较高的用户活跃度相关，但调研结果显示，API调用量和用户活跃度并不总是成正比。因此，在发展过程中需要平衡API调用量和用户活跃度，不能仅仅依赖于调用量来评估AI应用的影响力和价值。

API提供方服务模式的影响

- API提供方的服务模式对API调用量和用户活跃度之间的关系产生影响。API提供方可以直接面向终端用户、企业客户或开发者，或同时拥有这两种模式。不同的服务模式会导致API调用量和用户活跃度之间的差异。因此，API提供方需要根据具体情况制定适合的服务模式，以提供优质的服务，吸引和留住用户。
- API调用与用户活跃度的平衡发展是AI团队业务发展的重要因素。不仅要关注API调用量的增长，还要注重提供优质的服务，吸引用户并保持用户的活跃度。同时，根据不同的服务模式制定相应的策略，以实现平衡发展。

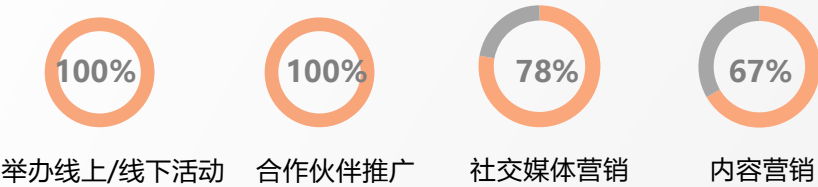
大模型垂直应用行业部署与应用成熟度



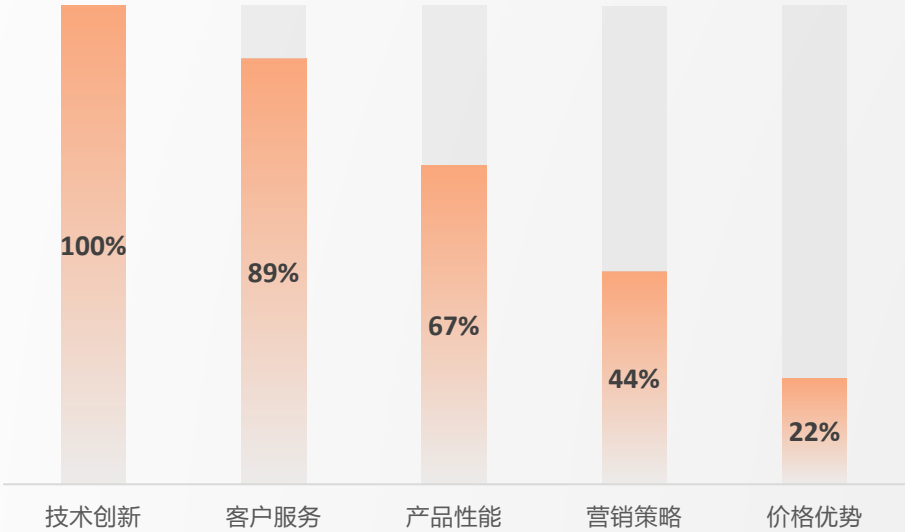
Q：据您了解，贵公司大模型商业应用和部署中，应用于以下哪些行业？备注：渗透率=选择该行业部署样本数量/参与调研的样本总数
Q2：据您了解，贵公司大模型在该领域的最成熟的应用包括以下哪些？备注：成熟度= 该行业选择占比超过50%的应用数量/该行业常见应用数量

客户拓展策略与成功要素

Q: 据您了解, 贵公司在客户拓展方面采取了哪些策略?



Q: 据您了解, 您认为在大模型商业化过程中, 哪些因素是成功的关键要素?



人

大模型可以应用于多种场景和领域, 但并不意味着它可以面向所有类型的客户。不同的客户有不同的需求、预算、能力和偏好, 因此大模型商业化需要明确目标客户群体, 并根据其特点提供个性化服务。

例如, OpenAI在公开GPT-3论文后, 开放了模型的API申请通道, 鼓励研究者、开发者、企业从业者研究“好玩”的GPT-3应用, 以此促动大模型的产业场景发展。OpenAI通过设置不同等级的API访问权限和收费标准, 区分了不同层次的客户, 并根据其反馈和需求进行持续优化。

货

大模型的核心竞争力在于其技术优势和应用价值, 因此在营销策略中, 需要突出创新性、性能、效率、通用性等特点, 以及其在不同场景和领域中带来的效果提升、成本降低、体验优化等价值。

例如: 华为在推出盘古NLP后, 通过发表白皮书、举办开发者大会、参与国际评测等方式, 展示了盘古NLP的技术特色和优势, 如超大规模参数、海量数据训练等。华为云还通过案例分析、数据对比、用户评价等方式, 展示了盘古NLP在能源、零售、金融、工业、医疗等领域中的应用效果和价值。

场

大模型作为一种新型高端技术, 需要选择合适的营销渠道和形式, 以便有效地传达信息和影响力。据与调人士, 9家公司都选择了线上线下活动和合作伙伴推广的方式进行营销。线上活动包括发布白皮书、举办技术分享会、参与国际竞赛、开放社区等; 线下活动包括举办技术论坛、参与行业展会、组织开发者大赛等。合作伙伴推广包括与行业领先企业或机构建立合作关系, 共同探索大模型的应用场景和商业模式。

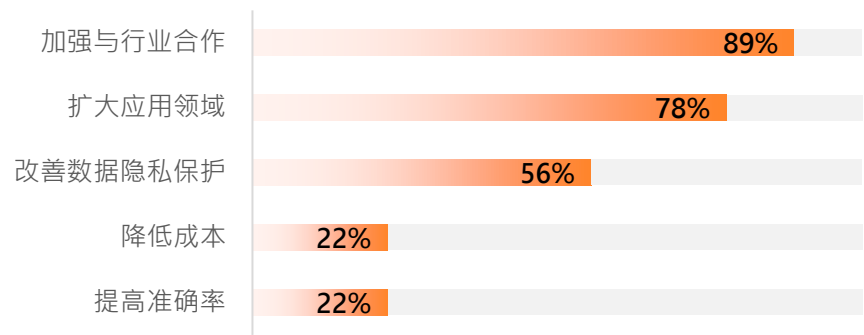


6

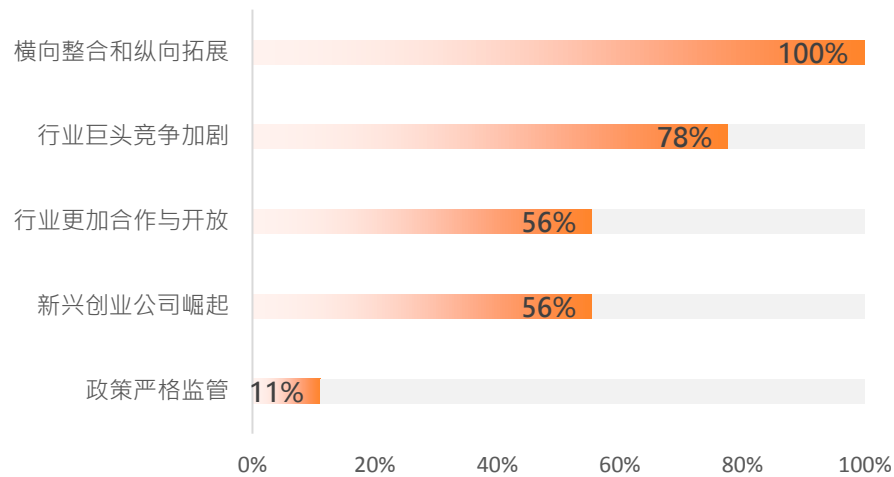
中国AI领域的未来： 整合，竞争，开放性与创新

行业合作与应用扩展优先，整合与竞争并行，创新不止

Q：据您了解，贵公司未来1年大模型发展的目标是？



Q：未来2年，关于中国AI大模型行业的竞争格局，您认同以下哪些？



- AI应用有广阔的发展空间，行业合作扩大影响力和覆盖面价值。AI不再仅仅是科技领域的专利，其跨行业的应用潜力正在逐步得到挖掘。各行业积极寻求与AI公司的合作，以此驱动行业的数字化和智能化转型。对于AI公司来说，加强与各行业的合作，不仅能够扩大他们的商业模式和营收来源，也有助于他们更好地理解并满足市场需求。
- 数据隐私保护的重视也在加强。随着数据安全和隐私问题日益突出，公众对于数据隐私保护的关注度也在逐步提高。面对这样的形势，近半数与调AI公司表示，他们将在未来一年推进数据隐私保护工作，旨在建立起更加健全的数据安全防护体系。
- 降低成本和提高模型准确率并不是这些公司的主要发展目标。我们认为可能因为在大模型领域，大模型发展初期，成本不是制约其发展的主要问题，公司们更加注重的是如何将AI技术融入更多的实际应用中，创造更大的商业价值

我们认为，未来两年的主要趋势预计将是行业的横向和纵向整合，竞争的加剧，以及行业的进一步开放与新兴创业公司的崛起。

- 行业的横向和纵向整合将是未来的一个主要趋势。AI公司将寻求在AI的不同子领域之间以及在AI产业链的上下游之间进行整合，以便提高效率，扩大规模，增强竞争力。随着技术的发展和市场的成熟，行业竞争预计将会加剧。这不仅表现在行业巨头之间的竞争，也表现在他们与新兴创业公司之间的竞争。
- 行业将会变得更加开放。有一半的受访者认为，新型的创业公司将会崛起。这预示着，在未来的AI领域，我们可能会看到更多富有创新的新企业出现。这些新兴创业公司的崛起，将为行业注入新的活力和创新，也将为消费者带来更多的选择和更好的体验。
- 关于行业监管，大部分受访者对此持乐观态度。我们认为，适度的政策监管有助于保护消费者权益，保障数据安全，促进行业健康发展。

风险提示

1. **样本代表性风险：**本报告基于产业问卷调研，由于样本量不足以覆盖所有中国AI科技公司的多样性和复杂性，被调研产业人士不能代表所在公司全貌，因此存在样本代表性风险。
2. **人工智能行业发展不及预期：**人工智能发展受到舆论道德争议，导致行业发展速度受限。
3. **商业模式仍不明朗：**AIGC应用处于起步阶段，商业模式尚未得到有效验证。
4. **法律风险：**对AIGC技术相关应用可能暗藏数据安全、著作权侵权、深度伪造、商业机密泄露、违法信息传播等风险。

分析师声明

本报告署名分析师在此声明：我们具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，本报告所表述的所有观点均准确地反映了我们对标的证券和发行人的个人看法。我们所得报酬的任何部分不曾与，不与，也将不会与本报告中的具体投资建议或观点有直接或间接联系。

一般声明

除非另有规定，本报告中的所有材料版权均属天风证券股份有限公司（已获中国证监会许可的证券投资咨询业务资格）及其附属机构（以下统称“天风证券”）。未经天风证券事先书面授权，不得以任何方式修改、发送或者复制本报告及其所包含的材料、内容。所有本报告中使用的商标、服务标识及标记均为天风证券的商标、服务标识及标记。

本报告是机密的，仅供我们的客户使用，天风证券不因收件人收到本报告而视其为天风证券的客户。本报告中的信息均来源于我们认为可靠的已公开资料，但天风证券对这些信息的准确性及完整性不作任何保证。本报告中的信息、意见等均仅供客户参考，不构成所述证券买卖的出价或征价邀请或要约。该等信息、意见并未考虑到获取本报告人员的具体投资目的、财务状况以及特定需求，在任何时候均不构成对任何人的个人推荐。客户应当对本报告中的信息和意见进行独立评估，并应同时考量各自的投资目的、财务状况和特定需求，必要时就法律、商业、财务、税收等方面咨询专家的意见。对依据或者使用本报告所造成的一切后果，天风证券及/或其关联人员均不承担任何法律责任。

本报告所载的意见、评估及预测仅为本报告出具日的观点和判断。该等意见、评估及预测无需通知即可随时更改。过往的表现亦不应作为日后表现的预示和担保。在不同时期，天风证券可能会发出与本报告所载意见、评估及预测不一致的研究报告。

天风证券的销售人员、交易人员以及其他专业人士可能会依据不同假设和标准、采用不同的分析方法而口头或书面发表与本报告意见及建议不一致的市场评论和/或交易观点。天风证券没有将此意见及建议向报告所有接收者进行更新的义务。天风证券的资产管理部门、自营部门以及其他投资业务部门可能独立做出与本报告中的意见或建议不一致的投资决策。

特别声明

在法律许可的情况下，天风证券可能会持有本报告中提及公司所发行的证券并进行交易，也可能为这些公司提供或争取提供投资银行、财务顾问和金融产品等各种金融服务。因此，投资者应当考虑到天风证券及/或其相关人员可能存在影响本报告观点客观性的潜在利益冲突，投资者请勿将本报告视为投资或其他决定的唯一

投资评级声明

类别	说明	评级	体系
股票投资评级	自报告日后的6个月内，相对同期沪深300指数的涨跌幅	买入	预期股价相对收益20%以上
		增持	预期股价相对收益10%-20%
		持有	预期股价相对收益-10%-10%
		卖出	预期股价相对收益-10%以下
行业投资评级	自报告日后的6个月内，相对同期沪深300指数的涨跌幅	强于大市	预期行业指数涨幅5%以上
		中性	预期行业指数涨幅-5%-5%
		弱于大市	预期行业指数涨幅-5%以下

THANKS