iResearch
艾瑞咨询

2024年中国AI基础数据服务研究报告

CONTENTS

目录

01 / AI基础数据服务行业概述

AI产业整体进展

多模态、长文本、大模型小型化成为热点研究方向

在过去几年里，大众已见识到GPT、 BERT等大语言模型在自然语言理解和生成方面的卓越能力。相比单一模态的大模型，多模态大模型能够提供更自然的人机交互方式，具备更全面和准确的认知能力，并在不同情境下表现出更高的鲁棒性，从而赋能更丰富和全面的AI应用。因此，多模态技术已成为诸多大模型厂商的研发重点。此外，长文本处理能力的提升，使大模型在理解和生成复杂文档方面表现更佳，能够更好地支持多主题和多步骤的推理任务；通过知识蒸馏、模型剪枝和混合精度训练等技术，大模型得以小型化，减少了计算资源需求，提高了推理效率，使大模型在资源受限设备上高效运行，提升了响应速度和用户体验，保护了用户的数据隐私。聚焦国内AI商业化市场，大模型商业化进程加速，API市场竞争激烈，价格战频现，但同时也反映出供应商间能力同质化的问题，亟需破局；另一方面，央国企凭借较好的数字化基础、丰富的数据资源及业务场景、相对充足的科技投入预算，成为现阶段国内大模型项目建设的主力军，推动了大模型在中国AI产业的商业化落地。

全球AI产品技术进展

多模态

- 概述：多模态大模型能够同时处理和理解包括文本、音频、图像和视频在内的多种数据类型，这使得它们能够提供更自然的人机交互方式，具备更全面和准确的认知能力， 并且在不同情境下表现出更高的鲁棒性，从而赋能更丰富和全面的AI应用
- 案例： 2024年5月， OpenAI推出GPT-4o，可对音频、视频和文本进行实时推理；

长文本

- 概述：长文本可支持模型理解和生成更复杂的文档、报告、小说等内容，能够更有效地进行知识管理和信息检索，提升了模型对于上下文理解的连贯性，进而更好地实现多主题、多步骤的复杂推理任务
- 案例： 2024年3月，月之暗面宣布旗下大模型产品Kimi开启200万字无损上下文内测，其后阿里、百度等大模型厂商均宣布相关大模型产品的长文本能力升级规划； 2024年4月， Google、 Meta等机构的研究人员先后提出Infini-attention、 Megalodon等无限

大模型小型化

- 概述：通过知识蒸馏、模型剪枝、混合精度训练等方法， “大模型小型化”相关技术可减少模型参数并降低计算资源需求， 提高推理效率，使大模型可在端边等资源受限的设备上高效运行，降低能耗，提升了响应速度和用户体验， 还增强了数据隐私保护，未来可能催生更多的创新型智能终端
- 案例： 2024年5月，微软表示Windows将附带40多个端侧AI模型，包括可用于搜索、

中国AI商业化落地进展

来源：艾瑞咨询研究院自主研究及绘制。

AP

为争夺大模型客户流量及背后云资源市场，24年上半年云厂商、大模型厂商等 相继调整API产品定价，低价甚至免费供应

价格战的积极意义

扩大客户量及使用频次，促使大模型技术在国内更快普及，加速创新型应用的诞生；促进供应商不断优化模型及计算架构，降低模型推理成本；竞争加速产业分层，较少社会整体资源消耗

价格战的另一面为大模型产品技术壁垒的薄弱

尽管大模型相关产品技术仍在迭代，但国内大模型尤其以API方式提供标准化大模型服务的各供应商的产品能力尚未形成较大代际差异；供应商需加速技术及产品差异化建设，获取足够的利润，产业才能健康、可持续的发展

央国企引领大模型项目建设

央国企对大模型的建设投入较多，与其有较好的数字化基础、丰富的数据资源 及业务场景、相对充足的科技投入预算相关

2024年上半年中国大模型相关项目中标统计

据智能超参数统计，2024年1-6月中国大模型相关项目中标数量达237个，前5个月披露的项目金额合计已过2023年；行业分布上，电信（47个）、能源（42个）位居1-6月的项目数量头两名，其次为教育、金融、政务等行业，各行业中的央国企均在积极推动大模型项目建设

来源：艾瑞咨询研究院自主研究及绘制。

数据、算法、算力是构建AI的三大要素

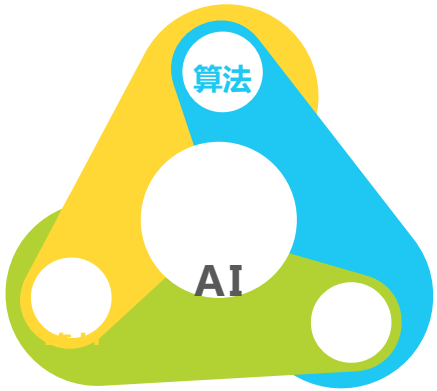
数据、算法、算力的协同促使现代AI技术实现了从理论到应用的飞跃

在人工智能领域，数据、算法和算力是构建AI系统的三大核心要素，三者的协同使现代AI技术实现了从理论到应用的飞跃。数据是AI的基础，大量高质量的数据不仅能够提高现有模型的准确率，还能促进模型的优化和创新。以ImageNet数据集为例，该数据集及相关挑战赛推动了计算机视觉算法的快速发展，2017年是挑战赛的最后一年，物体分类冠军的准确率在7年时间里从71.8%上升到97.3%。近年来，Transformer等预训练大模型在语言理解及生成等领域表现出色，大模型背后的Scaling Law（规模定律）进一步揭示了模型性能与数据量、算力之间的关系，强化了数据在提升AI表现中的关键作用。

构建AI系统的三大核心要素：数据、算法、算力

算法 是处理信息、提取特征、进行预测的逻辑框架

深度学习的兴起，CNN、Transformer等模型的迭代，极大地推动了图像识别、语义理解、文本生成等AI任务的进步



算力 支持算法处理庞大和复杂的数据集

GPU、TPU等AI芯片的发展，使得研究人员能够探索更深、更宽的网络结构，训练更强大的模型，并加速模型的推理速度。硬件的进步直接影响到AI模型的训练效率及规模化应用的可行性，从而不断拓展AI的边界

数据 是模型学习和适应不同任务的基石

高质量的数据能够帮助模型更好地理解现实世界，并做出更精准的预测；反之，即使是最先进的算法，也无法从劣质的数据中获得有效的洞察

来源：艾瑞咨询研究院自主研究及绘制。

高质量数据推动AI系统的发展进步

ImageNet数据集的成功，以及大模型的Scaling Law的发现，都证明着高质量 数据对于AI发展的巨大推动

ImageNet见证CV算法在大规模数据集上的性能提升

- 2009年6月，李飞飞团队完成ImageNet初始版本，共有1500万张图片，涵盖了 2.2 万个不同类别，这些图片筛选自近10亿张候选图片，并由来自167个国家的4.8万多 名全球贡献者进行了标注
- 2012年，由Alex Krizhevsky、 Ilya Sutskever和Geoffrey Hinton共同开发的

  AlexNet在挑战赛上以超过第二名10个百分点的成绩在夺冠，深度学习迎来学术探 索与工业应用的热潮
- 2017年是挑战赛的最后一年，物体分类冠军的准确率在7年时间里从71.8%上升到 97.3%，超越了人类的物体分类水平

Scaling Law进一步揭示数据对于提升模型性能的关键作用

- OpenAI研究团队于2020年发表的论文《Scaling laws for neural language

  models》中，系统地探讨了语言模型性能与模型大小、数据集大小和计算资源之间 的关系。研究发现，模型的性能（如损失函数值）与这些因素之间存在稳定的幂律 关系，即模型的性能会随着数据量、模型规模和计算量的增加而提升
- 现阶段，诸多大模型的研发仍在遵循Scaling Law的发展方向

  ① 今年2月，由ServiceNow、 Hugging Face 和 NVIDIA联合发布的用于代码生成 的StarCoder2，其数据集规模相比v1大7倍，实现了更准确的上下文感知预测

  ② 今年4月， Meta推出Llama3，其训练数据集超过15T token（是Llama2的7倍），可支持8K的上下文长度（是Llama2的2倍），在MMLU、GPQA、 HumanEval 等多项基准上成绩优异

来源：艾瑞咨询研究院自主研究及绘制。

d.book118.com/058041117015006111