

# 人工智能大语言模型 技术发展研究报告（2024 年）

中国软件评测中心  
（工业和信息化部软件与集成电路促进中心）

2024 年 6 月

人工智能作为引领新一轮科技产业革命的战略性和新质生产力重要驱动力，正在引发经济、社会、文化等领域的变革和重塑，2023年以来，以 ChatGPT、GPT-4 为代表的大模型技术的出台，因其强大的内容生成及多轮对话能力，引发全球新一轮人工智能创新热潮，随着大模型技术演进、产品迭代日新月异，成为科技产业发展强劲动能。本报告总结梳理大语言模型技术能力进展和应用情况，并对未来发展方向予以展望，以期为产业界提供参考。

由于编者水平所限，不妥之处，请批评指正。

## 目录

<b>第一章 大语言模型发展基石</b> .....	1
(一) 软硬协同持续推动大模型能力提升 .....	1
1.大模型发展对算力需求成井喷式增长 .....	1
2.AI芯片自研和算力优化成为应对算力需求的重要手段 .....	2
3.计算、存储、网络协同支持大模型训练 .....	3
4.深度学习框架是大模型研发训练的关键支撑 .....	5
5.大规模算力集群的创新应用与突破 .....	6
(二) 数据丰富度与质量塑造大模型知识深度与广度 .....	7
1.大模型对数据数量、质量提出新要求 .....	7
2.产业各方加快构建高质量丰富数据集 .....	11
(三) 算法优化与创新推动大模型能力升级 .....	14
1.多阶段对齐促进大模型更符合人类价值观 .....	14
2.运用知识增强提升模型准确性 .....	15
<b>第二章 大语言模型发展现状</b> .....	16
(一) 模型训练推理效率及性能明显提升 .....	17
(二) 围绕中文生成与推理能力构筑比较优势 .....	18
(三) 模型应用生态更加丰富多样 .....	18
(四) 海量数据处理基础能力不断增强 .....	19
(五) 采用多模型结合的路线加速应用落地 .....	20
<b>第三章 大语言模型的核心能力进阶</b> .....	22
(一) 深层语境分析与知识融合强化语言理解应用 .....	22
(二) 精确内容生成与增强搜索的融合 .....	23

(三) 符号逻辑与神经网络的融合提升 .....	25
(四) 上下文记忆能力的增强 .....	26
(五) 更为可靠的内容安全与智能应答机制 .....	27
<b>第四章 大语言模型创新应用形态——智能体 .....</b>	<b>28</b>
(一) 智能体 (AI Agent) .....	28
1. 智能体正成为大模型重要研发方向 .....	28
2. 大模型能力为 AI Agent 带来全面能力提升 .....	29
(二) 典型 AI Agent 案例 .....	32
1. RoboAgent: 通用机器人智能体的开创性进步 .....	32
2. Coze: 优秀的创新型 AI Agent 平台 .....	33
3. Auto-GPT: 推动自主 AI 项目完成的新范例 .....	34
4. Amazon Bedrock Agents: 企业级 AI 应用的加速器 .....	34
5. 文心智能体平台: 革命性的零代码智能体构建平台 .....	35
6. 腾讯元器: AI Agent 的智慧化体验 .....	35
7. NVIDIA Voyager: 引导学习的 Minecraft 智能体 .....	36
8. MetaGPT: 多智能体协作的元编程平台 .....	36
<b>第五章 大语言模型应用发展趋势 .....</b>	<b>37</b>
(一) 大模型将更加注重多模态数据融合 .....	37
(二) 大模型将提升自适应和迁移学习能力 .....	39
(三) 采用可解释性算法提高模型透明度 .....	40
(四) 垂直大模型产品研发需结合行业深度定制 .....	41
(五) 大模型发展需妥善处理隐私保护与数据安全问题 .....	43

## 第一章 大语言模型发展基石

### （一）软硬协同持续推动大模型能力提升

#### 1. 大模型发展对算力需求成井喷式增长

大规模的训练和推理需要强大的高性能算力供应，高端 AI 芯片是大模型高效训练和应用落地的核心，是决定大模型发展能力高低的关键。人工智能大模型参数规模和训练数据量巨大，需千卡以上 AI 芯片构成的服务器集群支撑，据测算，在 10 天内训练 1000 亿参数规模、1PB 训练数据集，约需 1.08w 个英伟达 A100 GPU，因大模型对高端 AI 芯片需求激增及高端芯片进口供应受限，英伟达等高端芯片已供不应求。据《金融时报》估算，我国企业对英伟达 A800、H800 两款 GPU 产品的需求达 50 亿美元。

GPT-3 的训练使用了 128 台英伟达 A100 服务器（练 34 天）对应 640P 算力，而 GPT-4 的训练使用了 3125 台英伟达 A100 服务器（练 90—100 天）对应 15625P 算力。GPT-4 模型参数规模为 1.9 万亿，约为 GPT-3 的 10 倍，其用于训练的 GPU 数量增加了近 24 倍（且不考虑模型训练时间的增长）而目前正在开发的 GPT-5 模型预计参数量也将是 T-4 模型的 10 倍以上，达到 10 万亿级别，这将极大地提升大模型训练的算力需求。同时，各应用单位、科研院所科技企业的自研模型需求逐步增长，据工业和信息化部赛迪研究院发布的研究报告预测，到 2024 年年底我国将有 5%—8% 的企业大

模型参数从千亿级跃升至万亿级，算力需求增速会达到320%。

此外，未来在 AI 算力基础设施领域，将有越来越多的厂商采用定制化算力解决方案。在摩尔定律放缓的大背景之下，以往依靠摩尔定律推动着性能效益提升的途径越来越难以以为继，要想得到最佳的计算性能，必须依靠针对特定应用和数据集合的体系架构。特别是在 AI 大模型领域，不同厂商均有着不同的差异化需求，越来越多公司发现，一体适用的解决方案不再能满足其计算需求。为把每一颗芯片的性能、效率都发挥到极致，做到最佳优化，需要根据算法模型、工作负载等进行针对性优化。

## 2.AI 芯片自研和算力优化成为应对算力需求的重要手段

算力芯片是大模型的算力“发动机”，拥有算力资源的企业具备更强的竞争力，强大的算力资源可以加速模型训练、提升市场响应速度，强力支撑更复杂、更深层次的模型训练，从而提高模型的预测精度和整体性能。

在大模型的高算力需求推动下，大厂加强 AI 芯片研发力度，持续优化大语言模型所用的 **transformer** 架构。如，谷歌为其最新款的 Pixel 手机装上了自研 Tensor G3 芯片，让用户可以在手机端解锁生成式 AI 应用。微软宣布推出两款自研芯片 Maia100 和 Cobalt100。Maia100 用于加速 AI 计算任务，帮助人工智能系统更快处理执行识别语音和图像等任务。