CSTC中国评测

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技术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

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模型的参数规模为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。