

高质量大模型基础设施研究报告

(2024 年)

中国信息通信研究院人工智能研究所

2025年1月

版权声明

本报告版权属于中国信息通信研究院，并受法律保护。转载、摘编或利用其它方式使用本报告文字或者观点的，应注明“来源：中国信息通信研究院”。违反上述声明者，本院将追究其相关法律责任。

前言

随着大模型技术的飞速发展，模型参数量急剧增长，模型能力持续增强，智能应用百花齐放。基础设施的可用性决定了大模型研发及服务的效率，大模型服务的可用性又决定了智能应用的服务质量。在此背景下，高质量大模型基础设施成为推动大模型应用落地的关键要素。

目前，大模型基础设施普遍面临可用性低、稳定性差等问题，亟需从计算、网络、存储、软件和运维等多层面协同优化。在同等计算资源条件下，通过多系统协同优化提升基础设施可用性，不仅可以提高大模型开发效率和服务能力，还能有效降低应用成本，加速大模型规模化落地。

本报告聚焦大模型基础设施的五大核心能力领域：计算、存储、网络、开发工具链和运维管理，系统梳理了大模型发展对基础设施提出的新需求，剖析基础设施发展的关键技术，并提出体系化评价指标。同时，通过分析业界典型实践案例，为企业建设高质量大模型基础设施提供参考。

展望未来，大模型基础设施将与大模型一起迭代升级，并为大模型的规模化应用提供有力支撑。本报告力求为相关领域的研究与实践提供参考，但难免有不足之处，恳请各位专家和读者不吝指正。

目 录

一、 大模型基础设施概述	1
（一） 大模型基础设施概念与特性	1
（二） 大模型基础设施现状	4
二、 大模型基础设施挑战	6
（一） 计算资源分配粗放，利用率低成为新难题	7
（二） 海量数据处理低效，数据存储成为新瓶颈	8
（三） 并行计算规模攀升，网络通信成为新阻碍	10
（四） 模型参数急剧增长，开发效率成为新约束	11
（五） 基础设施故障率高，运维能力成为新挑战	14
三、 大模型基础设施关键技术	15
（一） 高效算力管理调度技术	15
（二） 高性能大模型存储技术	16
（三） 高通量大规模网络技术	18
（四） 高效能大模型开发技术	20
（五） 高容错大模型运维技术	22
四、 高质量大模型基础设施评价指标	23
（一） 指标体系	23
（二） 指标定义	25
五、 高质量大模型基础设施典型实践	27
（一） 案例一：Meta 大模型基础设施实践	27
（二） 案例二：蚂蚁集团大模型基础设施实践	29
（三） 案例三：某科技公司大模型基础设施实践	31
六、 总结与展望	33
附录 高质量大模型基础设施规划建议	35

图 目 录

图 1	大模型基础设施架构图	1
图 2	大模型基础设施能力矩阵	2
图 3	大模型全生命周期对大模型基础设施的关键需求	7
图 4	大模型基础设施网络互联	20
图 5	高质量大模型基础设施评价体系	24
图 6	Meta AI 集群系统框架图	28
图 7	蚂蚁大模型基础设施架构	30

表 目 录

表 1	大模型基础设施技术能力评价指标	25
表 2	大模型基础设施性能评价指标	26

一、大模型基础设施概述

大模型技术作为人工智能领域的突破性进展，正迅速推动各行各业的智能化转型。随着参数量的增长，大模型展现出强大的理解能力和复杂数据处理能力，在金融、医疗、政务等行业的应用日益广泛。然而，参数量的增加也给大模型落地带来了巨大的挑战，提高大模型基础设施能力，满足大模型全生命周期对基础设施的新需求成为首要任务。

（一）大模型基础设施概念与特性

大模型基础设施是指支持大规模人工智能模型（如大语言模型、多模态大模型等）训练、部署和应用的硬件和软件资源的集合，它包括高性能计算、海量数据存储、高速网络连接资源，以及相应的软件框架和工具链，旨在为大模型的开发、训练和推理提供高效、可靠和可扩展的支撑环境。大模型基础设施整体架构如图 1 所示。



来源：中国信息通信研究院

图 1 大模型基础设施架构图