人工智能技术与应用前沿

Frontiers of Artificial Intelligence Technology and Applications

包义明/BAO Yiming，林阳/LIN Yang，屠要峰/TU Yaofeng
（中兴通讯股份有限公司，中国 深圳 518057）
(ZTE Corporation, Shenzhen 518057, China)

DOI：10.12142/ZTETJ.202504010
网络出版地址：http://kns.cnki.net/kcms/detail/34.1228.TN.20250711.1947.004.html
网络出版日期：2025-07-14
收稿日期：2025-05-10

摘要：人工智能（AI）正以前所未有的深度和广度重塑人类社会的生产与生活方式。在AI技术层面，大模型的架构创新、研发范式变革、多模态大模型算法以及大模型训推基础设施不断突破。同时，AI工程化应用显著加速，端侧AI和具身智能等新型产品与应用取得了重要进展。DeepSeek通过在模型算法和工程优化上的系统级创新，为资源受限环境下探索通用人工智能开辟了全新路径。未来，人工智能的核心竞争力将取决于：AI算法及模型架构的持续创新，AI基础设施对算力、存储、网络等资源的高效综合调度能力，AI应用的快速迭代与场景化落地能力。以大模型技术为主线，系统分析梳理了大模型架构与研发范式、AI基础设施、大模型应用技术等领域的最新进展与未来发展方向，为人工智能研究与实践提供参考。

关键词：大模型技术；大模型训推基础设施；强化学习；思维链；智能体

Abstract: Artificial intelligence (AI) is reshaping the production modes and lifestyles of human society with unprecedented depth and breadth. At the AI technology level, there are continuous breakthroughs in the architectural innovation of large models, changes in R&D paradigms, multimodal large model algorithms, and large model training and inference infrastructure. At the same time, the engineering application of AI has been significantly accelerated, and new products and applications such as end-side AI and embodied intelligence have made important progress. DeepSeek has opened up a new path for exploring general artificial intelligence in resource-constrained environments through system-level innovations in model algorithms and engineering optimization. In the future, the core competitiveness of artificial intelligence will depend on: the continuous innovation of AI algorithms and model architectures, the efficient and comprehensive scheduling capabilities of AI infrastructure for computing power, storage, network and other resources, and the rapid iteration and scenario-based landing capabilities of AI applications. This article takes large model technology as the main line, systematically analyzes and sorts out the latest progress and future development directions in the fields of large model architecture and R&D paradigms, AI infrastructure, and large model application technology, and provides a reference for artificial intelligence research and practice.

Keywords: large model technology; large model training and inference infrastructure; reinforcement learning; chain of thought; AI agent

引用格式：包义明, 林阳, 屠要峰. 人工智能技术与应用前沿 [J]. 中兴通讯技术, 2025, 31(4): 67-74. DOI: 10.12142/ZTETJ.202504010
Citation: BAO Y M, LIN Y, TU Y F. Frontiers of artificial intelligence technology and applications [J]. ZTE technology journal, 2025, 31(4): 64-74. DOI: 10.12142/ZTETJ.202504010

近年来，主流大模型在技术路径上呈现出显著分化的趋势。OpenAI的GPT-4系列凭借Transformer架构的深度优化在自然语言处理任务中确立了标杆地位。Google的Gemini通过多模态融合技术实现了文本、图像和语音的统一建模。Meta的Llama3则以开源策略推动了社区生态的繁荣发展。然而，这些模型普遍依赖海量参数规模和巨额算力投入来追求性能提升，逐渐形成了“大力出奇迹”的行业共识。

2024年底，DeepSeek通过在大模型算法和工程优化方面进行系统级创新，打破了AI在扩展定律下的预期天花板，为大模型的高效训练与推理提供了有效参考，并为在受限资源下探索通用人工智能开辟了新的道路。这标志着技术竞争焦点已转向架构创新、算法效率与场景适配的深度融合。

1 大模型技术

如表1所示，以Transformer架构为核心的人工智能技术前沿创新主要集中在3个方向：1）大模型架构创新，主要对大模型中的关键组件进行优化；2）大模型研发范式创新，以数据工程、大模型后训练、强化学习等手段，推动大模型逼近通用人工智能（AGI）；3）多模态大模型，大模型的能力从自然语言建模扩展到图片、视频、音频等多种模态的建模、理解和生成，向“世界模型”迈进。通过这些创新，Transformer及其变体模型不仅在性能上实现持续突破，还在

表1 近年来大模型技术前沿创新

技术创新类型	细分类别	方法/论文
模型架构创新	Attention层优化	MHA[1]、GQA、Flash Attention[2]、Selective Attention[3]、MLA[4]、NSA[5]
	MLP层优化	Gshard、DeepSeek-MoE[6]、HMoE[7]
	其他算子优化	DyT[8]、Scalable Softmax、RoPE[9]、YaRN[10]、Self-Extend、MRoPE[11]
	非Transformer架构创新	Mamba、Nemotron-H[12]、Hunyuan-T1[13]、RWKV
研发范式创新	强化学习与思维链	InstructGPT、PPO、CoT[14]、PRM[15]、OpenAI o1[16]、GRPO[17]、DeepSeek-R1[18]、DAPO[19]、VAPO[20]
多模态大模型	多模态架构创新	Flamingo[21]、DeepSeek-VL2[22]、Chameleon[23]、Emu3[24]、Stable Diffusion[25]
	多模态技术创新	ViT[26]、Q-Former[27]、DiT[28]

CoT：思维链
DAPO：解耦剪切与动态采样策略优化
DiT：扩散转换器
DyT：动态双曲正切函数
GQA：分组查询注意力
GRPO：群体相对策略优化
HMoE：异构混合专家
MHA：多头注意力
MLA：多头潜在注意力
MLP：多层感知机
MoE：混合专家
MRoPE：多模态旋转位置编码
NSA：原生稀疏注意力
PPO：近端策略优化
PRM：过程监督奖励模型
RoPE：旋转位置编码
RWKV：响应加权键值
VAPO：基于价值的增强近端策略优化
ViT：视觉转换器
YaRN：另一种RoPE扩展方法

效率、成本和适用性方面实现了显著优化，为人工智能的进一步发展奠定了坚实基础。

1.1 大模型架构创新

1）注意力机制创新

Transformer架构之所以成为当前人工智能革命的核心驱动力，其根本原因在于注意力机制对序列建模的范式革新。自注意力层通过动态计算Token间相关性实现了全局依赖捕获，而多头注意力（MHA）[1]则进一步提升了模型对不同维度和距离的交互模式学习能力。然而，这种强大的建模能力也带来了较高的计算复杂度$O(N^2)$和显存占用问题。为了降低计算资源的消耗，同时保持模型的性能，分组查询注意力（GQA）通过对查询（Query）进行分组，使组内查询矩阵共享键值（Key, Value）矩阵，从而有效减少计算量。Flash Attention[2]则提出分块计算与重计算策略，将内存访问量由二次方降至线性级。选择性注意力[3]通过动态筛选和压缩来减少计算量。DeepSeek提出的多头潜在注意力（MLA）[4]则颠覆性地将键值矩阵进行低秩压缩，在无损的前提下将内存开销降低超50%。

2）MLP层创新

在Transformer架构中，由MLP构造的前馈网络（FFN）通常占据整个模型50%以上的参数量。因此，针对MLP层的优化和创新成为提升模型效率的关键方向，其中稀疏化技术尤为重要。例如，混合专家（MoE）门控机制通过选择性地激活部分FFN层的参数，在降低推理复杂度的同时，实现对输入Token的“分而治之”。2020年，Google提出的Gshard首次将MoE机制引入Transformer架构。然而Gshard中的MoE存在专家负载不均衡和专家能力不足的问题。为了解决这些问题，DeepSeek[6]提出了细分小专家的方法，并在图形处理器（GPU）设备粒度上设置Token路由上限，实现了细粒度的负载均衡。DeepSeek和HMoE[7]等模型还引入异构差异化专家，通过结合细分领域与通用领域知识，使MoE模型能够更好地满足不同任务需求。这种异构专家的设计显著提升了大模型的泛化能力和领域适应性。

3）其余算子创新优化

除了Attention层与MLP层，Transformer架构中其他关键算子的优化也取得了显著进展。在归一化层，何凯明等提出的动态反切归一化（DyT）[8]使用动态缩放的Tanh函数替代传统的LayerNorm和RMSNorm，大幅提升了计算效率。这一创新标志着算子优化从“统计学驱动”向“几何学驱动”的转变。在位置编码方面，苏剑林等提出的旋转位置编码（RoPE）[9]成为相对位置编码的重要里程碑，显著提升了大模型的外推能力，支持了长序列建模的实现。Qwen2.5-VL[11]等将RoPE算子从纯文本建模推广到多模态领域，使模型在长视频理解等任务中表现卓越。

4）非Transformer架构创新

近年来，具有线性复杂度的大模型架构不断涌现。例如，Mamba架构基于选择性状态空间模型（Selective SSM），能够根据输入内容动态调整模型参数，对敏感内容进行重点建模，同时选择性地遗忘无效信息。这一特性使Mamba在处理长序列任务时表现出色。此外，Mamba-Transformer混合架构也逐步成熟。英伟达的Nemotron-H[12]和腾讯的Hunyuan-T1[13]充分结合了Mamba处理长序列的能力与Transformer的局部建模能力，实现了速度和精度的高度平衡。而接收加权键值（RWKV）架构在Transformer中融合RNN的能力，通过动态状态演化大幅度降低了大模型的边缘部署成本。

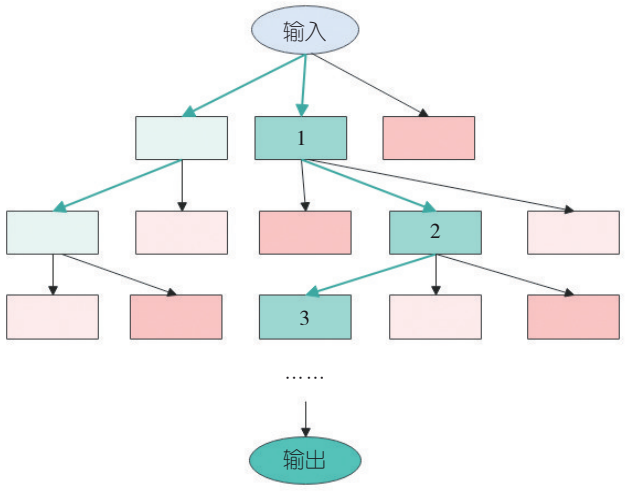
1.2 强化学习与思维链

大模型的训练旨在提升人工智能对世界的理解能力。在预训练（PT）-监督微调（SFT）-人类反馈强化学习

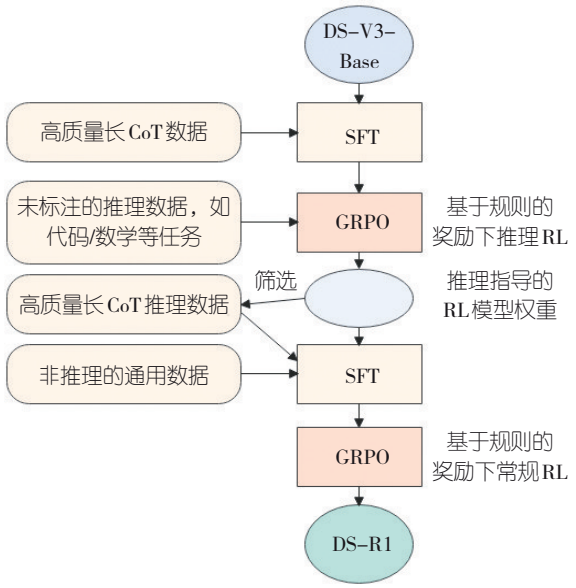
（RLHF）的大模型研发范式下，基于Next Token Prediction的PT和SFT本质上是一种“行为克隆”，其所学知识来源于人类语言，因此模型的能力受限于标注或无标注的训练数据。然而，近年来互联网中可用的高质量训练数据几乎已被消耗殆尽，大模型的能力也因此遇到了瓶颈。RLHF结合近端策略优化（PPO）算法在生成式预训练转换器（GPT）系列模型中的成功应用，为大规模人类反馈学习注入了新知识，同时显著提升了大模型的安全性。然而，RLHF的训练成本很高，因为人类对齐所需要的大量奖励信号需要人工标注。此外，基于这些奖励的大模型后训练也耗资巨大。

为突破大模型训练效率瓶颈，使其通过自我学习获得超越人类的知识和能力，另一种更为通用的强化学习范式——自我博弈（Self-Play）迅速发展。这种范式催生了OpenAI o1和DeepSeek R1等具有自我纠偏和链式思考能力的大模型，推动了新一波人工智能革命浪潮。在强化学习的理论与实践中，智能体（Agent）通过执行动作（Action）从环境中获取奖惩信号（Reward），进而进行自我调整并更新状态（State）。在RLHF任务中，通常采用PPO算法训练大模型，这一过程需要构建4个模型，即Actor模型、Reference模型、Reward模型和Critic模型。其中，Actor和Critic模型需要实时训练与更新参数，Reference和Reward模型仅进行推理以提供参考信息。为了让大模型获得自我纠偏能力，以Google和OpenAI为代表的研究团队在过去两年研究出了一套较为成熟的强化学习范式。这一范式以PPO算法为基础，结合思维链（CoT）[14]与过程监督奖励模型（PRM）[15]等技术，使得大模型在训练和推理时具备了显式思考和自我纠偏的能力。

如图1（a）所示，CoT通过将复杂问题拆解为多步骤的中间推理链，引导模型显式展示“思考过程”，从而提升解决复杂任务的准确性。例如，在数学问题中，模型需先列出公式，再计算中间结果，最后推导出答案。PRM则对思维链的每一步推理结果进行评价和验证。这种密集的奖励反馈信号，显著提升了大模型在长思维链任务中的稳定性、安全性与准确性。尽管上述方法取得了良好的性能效果，但过程监督的密集计算，以及蒙特卡洛树搜索等思维链涉及技术的搜索广度，大幅增加了强化学习的成本。为了降低强化学习算法PPO的复杂度和资源消耗，DeepSeek提出了群体相对策略优化（GRPO）[17]，通过对组内多候选输出进行归一化奖励计算当前的相对优势，从而去除了Critic模型，其训练速度相较PPO提升30%，显存占用减少50%。此外，DeepSeek R1还设计了一种基于规则的奖励策略。结合答案准确性奖励和思考过程的格式约束奖励，在经过GRPO训练后，DeepSeek R1模型具备自我反思能力，出现了“Aha Moment”训练流程，如图1（b）所示。近期，字节跳动也提出了解耦剪切与动态采样策略优化（DAPO）[19]与基于价值的增强近端策略优化（VAPO）[20]算法，为其大模型Seek-Thinking V1.5提供了更优的强化学习方案。其中，



（a）思维链式推理流程



（b）DeepSeek R1训练流程

CoT：思维链　DS：DeepSeek　GRPO：群体相对策略优化　RL：强化学习　SFT：监督微调

图1　CoT式推理与DeepSeek R1模型训练流程

VAPO是基于价值方法的SOTA方案，而DAPO则是无价值方法的SOTA方案。

1.3 大型多模态模型

早期的多模态模型主要聚焦于特定任务，通用性较差。基于Transformer的视觉语言预训练模型通过对比学习图像与文本特征，需要针对下游任务进行微调。多模态大语言模型（MLLM）将非文本输入和输出融入大语言模型（LLM），赋予LLM多模态信息处理能力。MLLM的典型架构通常包含5个组件，即模态编码器、输入映射器、LLM、输出映射器和模态生成器。Flamingo[21]是首个在视觉语言领域将LLM与视觉模型结合的MLLM。DeepSeek-VL2[22]利用MoE架构提升推理效率，由视觉编码器、视觉语言适配器和MoE语言模型组成。

尽管MLLM通过复用语言模型和其他模态模型降低了训练成本，但其跨模态能力较为受限。最近两年，统一多模态模型得到了快速发展。这类模型通过单一Transformer架构，统一处理多种模态数据，无需额外的对齐过程，通过预测下一个离散表示实现了跨模态的推理与生成。Chameleon[23]基于Transformer模型统一处理图像、文本、代码的离散Token。Emu3[24]的文本、图像和视频使用同一个离散Token空间。

MLLM在多模态学习中依赖模态编码、模态对齐和模态生成等多项核心技术。模态编码将不同模态的数据转换为统一特征表示，包括词嵌入模型、预训练语言模型、预训练音频模型、基于ViT[26]或卷积神经网络（CNN）的视觉模型等。模态对齐将不同模态的特征进行对齐，包括由线性层与非线性层交替组成的MLP，通过查询向量从视觉模型中提取视觉特征的Q-Former[27]等。模态生成是指特定模态的生成输出。扩散模型是视觉生成的主流方法。Stable Diffusion[25]是基于扩散模型的图像生成技术。Diffusion Transformer[28]是在Stable Diffusion的基础上，将主干U-Net替换成Transformer的图像生成模型。

MLLM通常采用分阶段训练策略，模态对齐负责弥合不同模态之间的差距；指令微调使模型理解和遵循多模态场景下的指令；偏好学习将模型的输出与人类的偏好对齐。多模态Prompt工程包括多模态CoT和多模态检索增强生成（RAG）[29]。其中，多模态CoT通过在输入中添加显式的思维链提示，引导模型逐步推理；多模态RAG将检索到的多种模态信息融入上下文，帮助模型全面理解任务背景并生成更优质的输出。

随着技术的快速迭代，MLLM展现出强大的跨模态处理能力，能够高效完成复杂的多模态任务。这些模型在图像生成、视频理解、音频分析等领域推动了下游任务性能的边界。

2 大模型训推基础设施

大模型训推基础设施通过“算力解耦”与“资源池化”破解了大模型成本困局。近期，DeepSeek创新众多AI Infra技术，大幅降低了上千亿参数模型的训练成本。未来10年，人工智能企业的核心竞争力将取决于AI Infra对算力、存储、网络的综合调度能力。

2.1 大模型高效率训练

随着参数规模的迅速飙升，大模型训练过程面临的性能瓶颈愈发显著，主要体现在内存墙、通信墙和存储墙3个方面。其中，内存墙尤为突出，主要受限于GPU显存容量。为突破上述性能瓶颈，表2总结了相关优化技术与解决思路。

1）大模型训练显存优化

大模型训练优化技术的发展与硬件（如显卡）的迭代相辅相成，其中分布式并行计算是解决GPU显存溢出问题的核心方法。Pytorch框架中的数据并行将一个批次中的样本分发到不同GPU上同时进行计算。Megatron-LM提出张量并行技术，将模型中间张量切分到多张GPU。流水线并行技术

表2 大模型训练效率优化技术

训练效率优化	优化技术或思路	方法/论文
显存优化	分布式并行训练	数据并行、张量并行、流水线并行[30-31]、序列并行[32-33]、专家并行[34]
	ZeRO技术	DeepSpeed ZeRO、CPU Offload
	其他显存优化技术	自动并行搜索[35]、混合精度训练[36]、梯度累积
通信优化	节点间GPU通信优化	InfiniBand、RoCE
	节点内GPU通信优化	PCIe、NVLink
	计算-通信重叠	Centauri[37]、DualPipe[38]
存储优化	高效文件I/O	DeepSeek-3FS[39]
	异步检查点I/O	DLRover[40]

CPU：中央处理器 GPU：图形处理器 I/O：输入/输出 PCIe：高速串行总线接口 ZeRO：零冗余

如Gpipe[30]、ZeroBubble[31]和DeepSeek DualPipe[38]等，将模型的不同层级按流水线方式分布到多GPU中执行。DeepSpeed推出的零冗余优化（ZeRO），将模型参数、优化器状态及张量梯度切分到多个GPU中存储。近年来，随着大模型训练的序列长度不断增长，序列并行成为大模型增强长下文建模能力的重要技术，以DeepSpeed Ulysses[32]和Ring Attention[33]两种序列并行（SP）技术的应用为基础，全球大模型得以支持128k甚至百万Token长度的序列建模。随着稀疏专家（MoE）架构模型的发展，专家并行技术被用于提升MoE模型的训练效率。为避免对上述多种并行方式的参数进行低效的人工调优，中兴通讯与北京大学联合研发了Galvatron[35]等自动并行优化技术，能够通过动态规划算法自动搜索最优并行参数。此外，面向显存优化的其他手段还包括FP8低精度训练[36]、梯度累积和参数量化等。

2）大模型训练通信优化

大模型分布式训练的效率通过GPU的“吞吐率（Throughput）”来衡量，而GPU之间的通信效率对训练效率起着决定性作用。在训练过程中，数据、模型参数以及中间结果需在不同GPU之间同步传递，通信延迟直接影响计算与通信的协同效率，从而可能导致资源闲置。为此，通信优化主要从两个层面展开：在硬件层面，多种高带宽组网协议推动着GPU集群通信效率的不断提升，例如以InfiniBand和基于融合以太网的远程直接内存访问（RoCE）为代表的节点间GPU通信协议，和以PCIe和NVLink为代表的节点内GPU通信协议；在软件框架层面，计算–通信重叠技术通过细粒度的算子拆分与编排来最大化通信效率，减少因数据传输导致的计算停滞时间。

3）大模型训练存储优化

在大模型训练中，存储系统扮演着关键角色：模型训练需要频繁地从存储中加载数据，同时需要实时将模型权重（检查点）保存到存储中，以避免因训练出错导致的算力浪费和研发进度延误。因此，高效的文件输入/输出（I/O）系统对于大模型训练至关重要。DeepSeek设计的3FS[39]系统，采用内存、固态硬盘（SSD）、分布式集群三级缓存结构，大幅提升了数据读取吞吐量，为大模型训练提供了高效的数据加载能力。为提升检查点读写速度，异步检查点技术如DLRover[40]等通过先将模型权重保存至内存，再在后续训练过程中异步写入磁盘，显著减少训练与存储操作的冲突，提升了存储和训练效率。

混合精度训练和量化是提升大模型训练效率的关键手段。混合精度训练通过结合单精度（FP32）和半精度（FP16）格式，在训练过程中动态调整数值精度，既减少了显存占用，又加快了计算速度。量化技术则进一步将模型参数和激活值从浮点数压缩为低位宽整数，有效降低模型大小和计算复杂度。

综上所述，通过显存、通信与存储优化技术的协同发展，大模型训练效率得以显著提升，同时推动了硬件与软件技术的进一步演化。这些优化手段为应对大模型训练中的性能瓶颈提供了系统性解决方案。

2.2 大模型高效推理

随着Agent、RAG等AI应用的快速发展，大模型的高效部署与推理已成为推动AI应用落地的关键需求。本节对比大模型高效推理技术，从计算效率、资源利用率等方面分析优化技术及其效果，详细总结见表3。

系统级优化通过传统网络和存储技术的改进优化推理流程。预计算与编码（PD）分离架构通过高性能网络将Prefill阶段和Decode阶段分开部署到异构集群，从而实现负载分离，显著提升推理吞吐性能。键值缓存（KV）Cache扩展利用存储设备实现对KV Cache的扩展，有效缓解内存压力，

表3 大模型推理优化技术

优化类型	技术	原理	效果	应用
系统优化	PD分离	推理的Prefill、Decode阶段分开部署	异构硬件部署，实现更高资源利用率	Kimi、DeepSeek
	多级KV缓存	显存、RAM、SSD、存储构成多级KV Cache存储	扩展KV多级缓存空间，支持更高吞吐率	Kimi、DeepSeek
模型结构	投机采样	用较小模型推理草稿，用大模型验证结果	有效降低大模型解码的次数	Medusa、GPT-4
	MTP	预训练n个独立Decode网络并行预测n个Token	缩短训练周期，提高推理效率	DeepSeek
	MLA	预训练低秩矩阵，实现KV的压缩存储	压缩历史KV的存储空间，提高端到端推理性能	DeepSeek
注意力稀疏化	MoBA	对输入序列进行分块，最关键的分块参与技术	以较少的存算资源实现接近原生注意力的精度	Kimi
	NSA	融合3种策略，选择最关键的KV参与注意力计算	极低的存算开销实现优于原生注意力的精度	DeepSeek
长思维链	隐藏空间推理	通过高维向量数学运算完成推理，将结果解码为可读形式	全新思维链范式	Coconut

KV：键值　MoBA：混合块注意力　NSA：原生稀疏注意力　RAM：随机存取存储器
MLA：多头潜在注意力　MTP：多Token预测　PD：预计算与编码　SSD：固态硬盘

同时进一步优化推理性能。

在模型结构方面，投机采样是一种并行推理技术，通过提前预测多个后续Token来提升推理效率。Meta和DeepSeek提出的多Token预测（MTP）[41]技术利用多个独立的Decoder输出网络并行预测多个后续Token，从而提升推理速度。DeepSeek的MLA[4]通过改进Transformer架构，有效压缩KV张量的存储空间。

注意力稀疏化技术旨在降低注意力机制的计算复杂度，通过捕捉推理过程中的稀疏性，仅对关键Token执行注意力计算，就能减少存算开销。Kimi提出的混合块注意力（MoBA）[42]算法，通过将注意力划分为若干块级区域，显著减少计算复杂度。DeepSeek提出的原生稀疏注意力（NSA）[5]策略，动态选择关键Token进行注意力计算，进一步优化推理性能。

长思维链推理优化关注在保证推理质量的同时减少资源开销。传统方法通常通过直接压缩推理步骤来降低计算成本，但这可能会影响推理效果。隐藏空间推理则通过潜在空间中的数学运算替代自然推理过程[43]，避免冗长的推理路径，成为长思维链优化的新范式。

综上所述，大模型高效推理需要从系统架构优化、模型轻量化、稀疏化技术以及推理范式创新等多方面入手，综合平衡推理性能、资源利用率与推理的可解释性，才能全面满足AI应用对高效推理的需求。

3 大模型应用技术

3.1 数据与大模型

数据是大模型能力的根基与驱动力[44]。在预训练阶段，海量高质量语料使模型得以掌握通用知识、语言模式和复杂语义结构；在推理阶段，结合检索增强生成（RAG）技术，动态获取外部知识为大模型提供实时、可信的信息补充，有效提升生成结果的准确性与时效性。

1）语料工程

大模型的预训练语料主要分为两类：一是由人类创作的原生数据，二是由模型生成的合成数据。随着模型参数规模的指数级扩展，原生高质量数据的增长速度远无法满足大模型训练的需求，合成数据正逐步成为下一代语料构建的关键。

一种主流的合成数据方式是结合知识蒸馏，从强大的教师模型中蒸馏知识用于训练轻量模型，实现性能迁移。在实际工程中，通常将合成数据生成与知识蒸馏联合应用：先利用大模型生成高质量训练样本，突破原生数据在规模和质量上的瓶颈；再通过蒸馏技术提升小模型性能，实现知识的高效迁移。例如，DeepSeek R1[18]通过动态温度参数调节，使3B模型性能逼近70B模型。OpenAI基于o1[16]生成合成语料训练Orion，并通过蒸馏实现对边缘设备的适配，形成从语料生成到端侧部署的闭环优化。

2）RAG

RAG系统通常包含4个阶段：预检索（Pre-Retrieval）、检索（Retrieval）、后处理（Post-Retrieval）和生成（Generation）。Pre-Retrieval阶段将检索源的数据组织为结构化知识库，支持文本、图像、表格等多模态数据。该阶段涉及数据清洗、去噪、切分（Chunking）与知识结构化处理。Retrieval阶段可采用多种检索方式。关键字检索通过倒排索引定位相关文档，利用BM25等算法计算文档得分。稠密向量检索通过查询向量和文档向量之间的距离衡量两者的语义相似性。学习型稀疏向量检索通过模型生成高维学习型稀疏向量，结合了关键词检索和稠密向量检索的优势。当前主流的检索方式为混合检索模式，即融合多种检索机制以提升准确率。Post-Retrieval对检索结果进行重排序、信息压缩（如摘要融合）等处理，以提升上下文信息密度，减少无效片段对生成质量的影响。在Generation阶段，大模型在融合的上下文信息基础上进行内容生成，提供更具事实性与语义丰富度的输出结果。

在实际应用中，RAG已广泛应用于企业知识问答、文档总结、垂直搜索等场景。企业数据通常呈现为文本、图像、表格、图表等多种模态，面对图文混排、多模态融合等复杂任务，多模态RAG成为突破关键。然而，多模态RAG仍面临多方面的挑战，包括异构数据的预处理与标准化、不同模态信息的协同存储与高效整合、跨模态语义融合与对齐、多模态混合检索结果的排序优化与语义一致性控制等。

3）向量数据库

向量是大模型处理文本、图像等多模态数据的关键，而向量数据库则是专为高效存储与检索高维向量数据（如文本、图像、音视频的嵌入表示）而设计，是AI时代的关键基础设施。向量数据库的核心价值在于：1）支撑多模态数据处理。向量数据库是实现跨模态语义对齐的关键组件，可通过语义匹配与知识融合高效处理多模态数据。2）RAG技术协同。向量数据库通过外置动态知识来降低模型幻觉，提升响应的实时性，同时避免模型参数膨胀与重复训练；3）私域知识管理与库内训推。向量数据库管理私有化知识，可以基于数据库技术框架实现领域大模型的训练和推理，在降低部署复杂度的同时确保数据安全与隐私。

目前主流向量数据库包括Weaviate、Milvus、Faiss、Chroma、Qdrant以及EBASE AI，其关键能力对比详见表4。

表4　主流向量数据库关键特性对比

特性	内置文本嵌入	索引类型	文本搜索	混合检索	元数据过滤	多租户	最大向量维度
Weaviate	√	HNSW	√	√	√	√	65 535
Milvus	×	FLAT、HNSW等	×	√	√	×	32 768
Faiss	×	FLAT、HNSW等	×	×	×	×	无限制
Chroma	×	HNSW	×	√	√	×	无限制
Qdrant	×	HNSW	×	√	√	√	无限制
EBASE AI	√	IVFELAT、HNSW等	√	√	√	√	32 768

AI：人工智能　FLAT：线性扫描索引　HNSW：分层可导航小世界　IVFELAT：倒排文件索引和线性扫描索引

EBASE AI是中兴通讯面向大模型时代自研的AI数据库系统，融合多模数据存储与一站式RAG能力。目前，EBASE AI已广泛应用于中兴通讯大模型系统，在产品问答、代码生成、文档分析等场景中实现快速落地，展示出卓越的能力。

3.2 端侧AI

近年来，在终端部署大模型服务于个人用户，已成为实现“个人智能”的主要路径，即端侧AI。中兴通讯端侧AI的统一技术架构如图2所示，通过在基础软硬件层、技术服务层、业务与应用层进行贯通及创新，推动中兴通讯在端侧AI领域的技术突破与产品落地，助力行业引领。

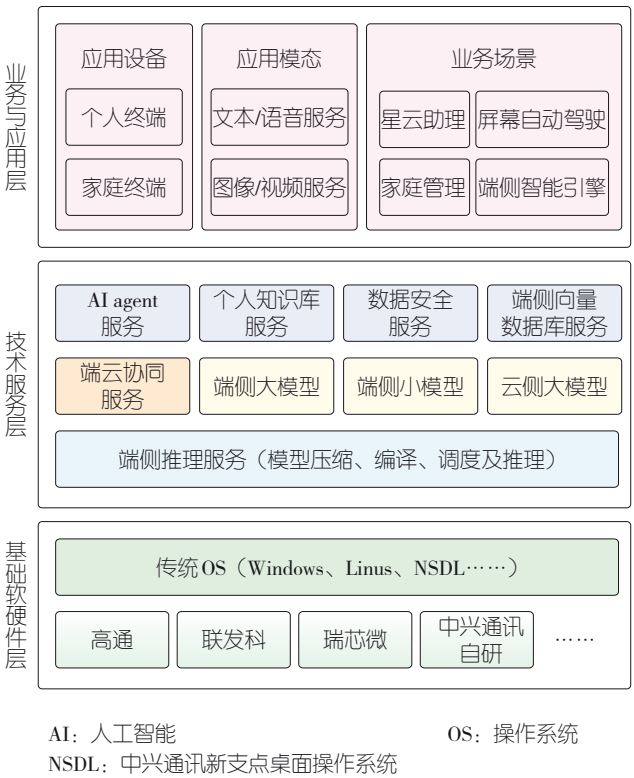


图2　中兴通讯端侧AI统一技术架构图

1）基础软硬件层

智能终端作为端侧AI的核心载体，其硬件架构与算力水平正迎来系统性升级。以高通骁龙8 Gen3、联发科天玑9400为代表的神经网络处理器（NPU）采用异构计算架构，大幅提升了大模型推理速度，同时将功耗降低35%。在硬件升级的同时，智能终端的软件生态也呈现深度协同的趋势。原生操作系统集成内核级AI能力，并通过意图框架解析用户指令，触发跨应用服务链。交互技术的创新尤为显著，基于苹果Ferret-UI和谷歌ScreenAI等面向端侧UI的视觉语言模型，支持屏幕元素的语义理解与自动化操作，使终端具备“以意图为中心”的类“端侧自动驾驶”能力，显著提升了智能交互体验。

2）技术服务层

在个人智能的愿景下，端侧AI服务需要更加贴合用户需求，兼具隐私性、情感化和个性化等多重特性。通过在端侧构建用户画像，并将用户画像数据存储到向量数据库，再结合RAG技术向通用大模型提供支持，是端侧AI优化的创新路径。

端侧向量数据库通过存储-计算-索引的协同设计，能够在端侧AI资源受限场景中实现高效检索与个性化服务。EBASE AI数据库端侧版本，通过对物联网（IoT）设备、智能手机和嵌入式系统等多样化硬件平台的无缝适配，已在资源受限的边缘与终端环境中展现出卓越能力。云边端协同通过异构资源的动态编排和调度，实现了计算负载与隐私安全的全局优化。边缘节点负责高实时性任务，云端专注于模型训练、更新与复杂任务推理，端侧聚焦个性化微调和轻量化的实时执行。

3）业务与应用层

端侧AI正加速向多形态终端渗透，形成了“核心智能终端+泛边缘设备”的立体化布局。在应用模式上，多模态融合成为端侧AI技术演进的主要路径。在业务场景上，个性化服务能力则成为技术竞争高地。例如，中兴通讯星云助

理为中兴通讯智能终端提供端侧自动驾驶等创新能力。这些基于端侧智能引擎的能力将进一步推动端侧AI的产业变革，加速多场景业务需求的落地，助力构建全面智能化时代。

3.3 智能体

大模型技术的飞跃推动了智能体（AI Agent）的快速发展[45-47]，逐步形成“模型突破→Agent应用涌现→场景扩展”的演进路径。各大厂商纷纷加码智能体技术，推出了一系列具有颠覆性的应用实践。ChatGPT催生了如Copilot这类AI增强助手，能够辅助人类完成写作、编程等任务，但尚不具备自主任务规划能力。而GPT-4强大的推理与规划能力则开启了“自主智能体”时代，衍生出AutoGPT等开源项目，以及全球首个全自动AI程序员Devin。随着OpenAI的o1和DeepSeek R1等具备复杂推理能力的大模型问世，Agent应用逐步具备了“思考+执行+验证”的完整任务闭环能力，为真正意义上的智能代理系统奠定了技术基础。

在应用场景复杂性日益提升的背景下，单一Agent已难以胜任多步推理与并发任务执行。因此，多智能体系统（MAS）应运而生。MAS通常由一个主控Agent协调多个专职Agent协作完成复杂任务。中国发布的Manus代表了智能体系统的最新发展，采用分工协作机制，实现了规划、执行、验证的闭环，突破了单体智能体难以覆盖的应用瓶颈。由于多Agent与Tools工具之间存在大量异步通信与上下文交互，统一的通信协议变得尤为关键。模块上下文协议（MCP）[48]已成为当前Agent系统中的主流通信标准，显著降低了系统开发与部署复杂度。近期，Google基于MCP推出扩展协议智能体到智能体（A2A）[49]，支持跨平台、跨厂商的智能体间安全协作，成为智能体社区的研究热点。

随着Agent系统日益智能化，其应用边界不断拓展，具身智能成为新一轮研究焦点。通过将智能体“具身化”，即赋予其感知现实世界并执行任务的能力，Agent技术正从虚拟场景走向真实环境，拓展出更广阔的应用空间。

3.4 具身智能

具身智能是指具备环境感知能力、目标规划能力以及实际执行能力的智能实体，通常表现为具有人类感知与行为能力的机器人系统[50]。随着宇树科技在四足机器人方面的突破，以及自动驾驶技术的商业化落地，具身智能逐步从实验室走向产业。具身智能被视为迈向AGI的关键路径之一，被英伟达的黄仁勋称为“AI的下一个浪潮”。

大模型的发展为具身智能带来多项核心能力突破，包括自然语言指令理解、视觉感知融合、多步动作规划与反馈式执行能力。传统虚拟Agent已开始“具身化”，通过集成硬件传感器和控制接口，将模型能力延伸至物理世界。当前主流的具身智能大模型有两类：一是以Google为代表的端到端具身大模型，二是以OpenAI为代表的分层决策模型。前者采用一个视觉-语言-动作（VLA）大模型完成感知、规划、执行全流程，如OpenVLA、TR-X模型基于Transformer架构统一处理视频、指令、动作序列[51]；后者通过分层模型处理不同任务，上层“大脑”基于预训练大模型负责感知、决策，中间“小脑”基于RL小模型实现动作规划，底层硬件“本体”负责精准执行[52]。智源研究院发布的大小脑协同系统RoboOS创新性地实现了“云端智能决策+终端实时执行”的协同框架，标志着具身智能正从单体智能向群体智能演进。

在产业应用方面，自动驾驶是目前具身智能落地程度最高的领域。百度“萝卜快跑”采用大模型分层控制架构，其安全性被证实优于人类驾驶员；而特斯拉的完全自动驾驶（FSD）系统则依赖端到端模型，直接将传感器输入映射为控制信号，实现驾驶自动化。

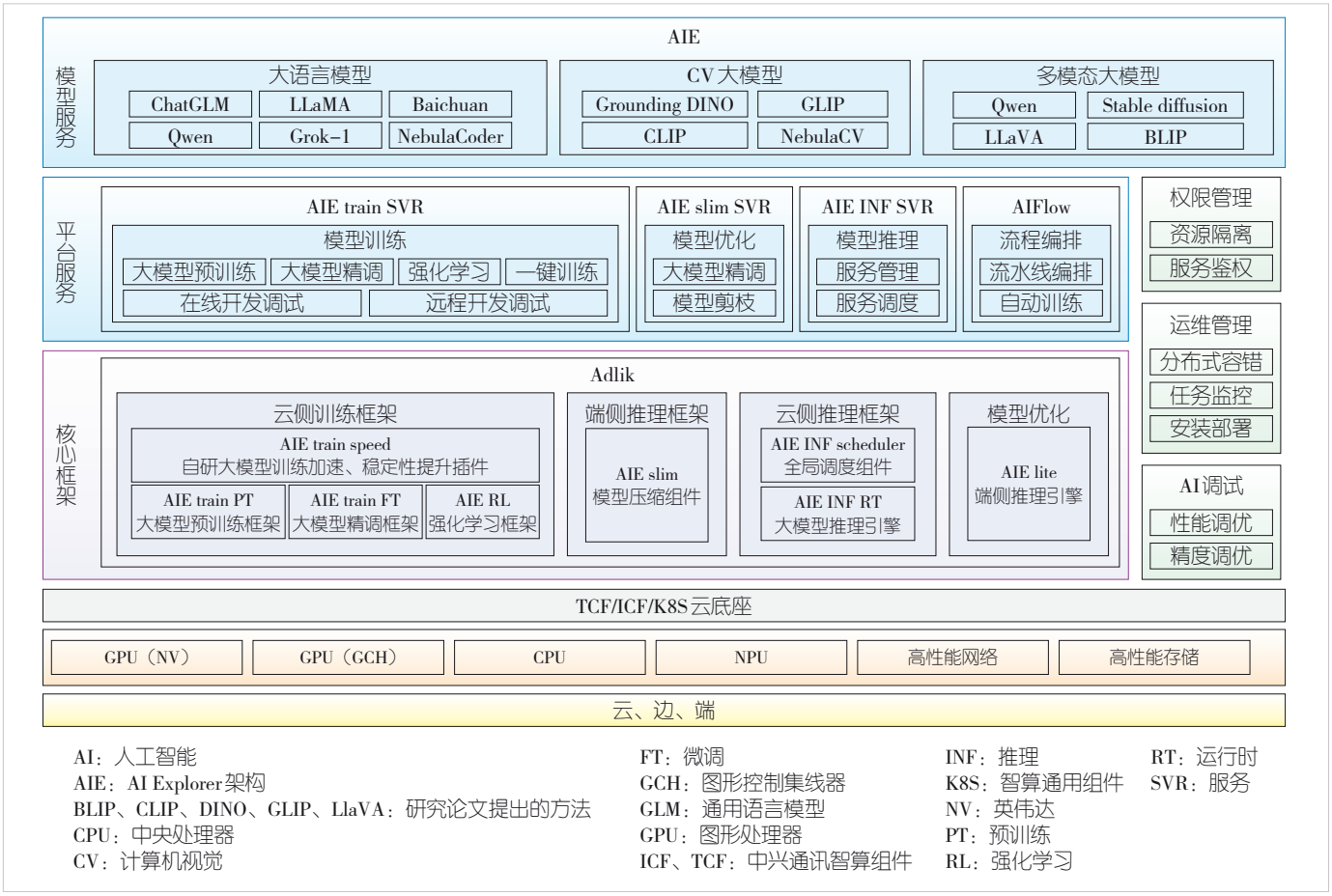
具身智能目前仍面临诸多挑战：高质量训练数据严重匮乏、仿真替代方案的保真度与通用性不足、开发工具与评估体系不成熟、真实部署复杂性高等。具身智能的发展亟需构建开放式生态体系，完善训练平台、通信协议和评估框架，以推动从“实验智能”走向“实用智能”的转变。

3.5 中兴通讯大模型及智算研发实践

中兴通讯在大模型技术创新方面持续深耕，勇于探索关键前沿方向，取得了显著成果。据SuperCLUE最新发布的《中文大模型基准测评2025年5月报告》显示，中兴通讯星云大模型Nebula Coder-V6在推理专项榜单中表现卓越，总分并列第一，同时在综合总榜中斩获银牌，展现出强大的模型能力与工程落地实力。围绕多模态大模型的产业化应用，中兴通讯已在工业制造、智慧教育、医疗健康等多个垂直场景中开展落地探索，未来将继续在强化学习、思维链推理、多模态融合和智算加速等核心技术方向加大投入，推动模型能力持续演进。

在算力基础设施方面，中兴通讯构建了完善的智算方案和AI平台，提供面向大模型研发的全栈AI Infra能力（如图3所示）。平台实现了数据、算力与算法的统一调度与集中管理，覆盖从数据采集、模型预训练、精调、强化学习到推理部署的完整开发流水线，显著提升了大模型研发的效率与稳定性。

在AI数据库与检索增强生成技术方面，中兴通讯基于



AI：人工智能
AIE：AI Explorer架构
BLIP、CLIP、DINO、GLIP、LlaVA：研究论文提出的方法
CPU：中央处理器
CV：计算机视觉
FT：微调
GCH：图形控制集线器
GLM：通用语言模型
GPU：图形处理器
ICF、TCF：中兴通讯智算组件
INF：推理
K8S：智算通用组件
NV：英伟达
PT：预训练
RL：强化学习
RT：运行时
SVR：服务

图3 中兴通讯的AI平台架构图

开源项目Dify[53]和RAGFlow[54]自主开发了高效检索系统NebulaRAG，如图4所示。该系统集成了多类开源生成式大模型、向量化嵌入模型与Rerank精排模型，实现语义检索与智能问答任务的高效闭环，构建面向行业知识库的多源智能交互平台。

在智能体应用方面，中兴通讯开源的Co-Sight超级智能体采用多智能体（Multi-Agent）架构，构建“数字团队”式协同体系。系统通过主管智能体统一调度各执行智能体，在有向无环图（DAG）机制支持下，实现并发任务优化与自适应容错，具备动态重规划与跨平台自动化执行能力，通过智能反思模型机制实现能力持续进化。在GAIA基准测评中，Co-Sight以72.72分位居开源框架榜首，已应用于行业研究、舆情分析、旅游规划等场景，大幅提升任务自动化水平与人效。

总体来看，中兴通讯在大模型算法架构、智算AI平台、智能检索系统与智能体应用等关键技术路径上持续发力，不仅实现了技术创新，更在多行业完成高质量落地。其在推动国产大模型生态建设、加速人工智能产业发展中，正发挥着日益重要的引领作用。

4 总结与展望

当前，大模型技术正沿着多元融合的路径演进，核心突破集中于算子优化、异构架构协同、强化学习，以及思维链驱动下的模型自学习能力增强，并朝着“理解-生成”一体化的多模态大模型架构不断演进。伴随着微调与后训练范式的日益普及，VeRL、Huggingface Transformers等开发框架持续演进，为大模型训练、部署与优化提供了更加灵活高效的支持体系。同时，智算服务器与异构算力资源的协同调度能力持续增强，为人工智能研发提供坚实的底层算力保障。在生态方面，以DeepSeek为代表的开源技术浪潮，推动了模型能力向更大范围普惠开放演进，促使AI技术生态从“闭源竞跑”转向“开源协同”。然而，开源体系下也暴露出潜在的滥用风险与安全漏洞，技术共享与主权控制之间的博弈加剧，亟需构建全球协同治理框架与标准体系，实现“安全可控”与“开放创新”的双重目标。

总体而言，大模型在复杂任务泛化能力、跨模态理解与

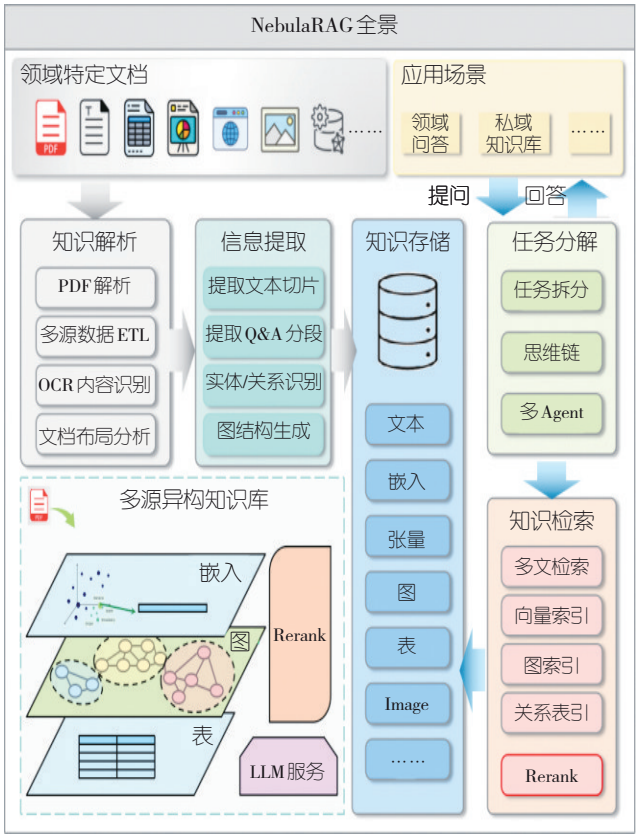


图4 NebulaRAG全景图

生成、推理路径稳定性、系统级集成能力等方面仍存在诸多挑战。硬件兼容性、部署效率与模型规模之间尚未实现理想平衡，多模型协同、Agent系统部署与安全治理体系尚待完善。

未来，人工智能的发展需在技术创新与风险防控之间实现动态平衡，在端侧智能化、具身智能、Agent协同等方向实现更深层次突破。跨学科协同与开源生态建设将成为AI下一阶段发展的主动力，助推其在通信网络、智能制造、医疗健康、教育服务等关键行业的深度融合与规模落地，真正实现从能力突破到价值创造的跃迁。

参考文献

[1] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [EB/OL]. [2025-04-05]. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
[2] DAO T. Flashattention-2: faster attention with better parallelism and work partitioning [EB/OL]. (2023-07-17)[2025-04-05]. https://arxiv.org/abs/2307.08691
[3] OHNSTON W A, DARK V J. Selective attention [J]. Annual review of psychology, 1986, 37: 43-75
[4] LIU A, FENG B, WANG B, et al. DeepSeek-V2: a strong, economical, and efficient mixture-of-experts language model [EB/OL]. (2024-05-07)[2025-04-05]. https://arxiv.org/abs/2405.04434
[5] YUAN J, GAO H, DAI D, et al. Native sparse attention: hardware-aligned and natively trainable sparse attention [EB/OL]. (2024-05-07)[2025-04-05]. https://arxiv.org/abs/2502.11089
[6] DAI D, DENG C, ZHAO C, et al. DeepSeekMoE: towards ultimate expert specialization in mixture-of-experts language models [EB/OL]. (2024-01-11)[2025-04-05]. https://arxiv.org/abs/2401.06066
[7] WANG A, SUN X, XIE R, et al. HMoE: heterogeneous mixture of experts for language modeling [EB/OL]. (2024-08-20) [2025-04-05]. https://arxiv.org/abs/2408.10681
[8] ZHU J, CHEN X, HE K, et al. Transformers without normalization [EB/OL]. (2025-03-13) [2025-04-05]. https://arxiv. org/abs/2503.10622
[9] SU J, AHMED M, LU Y, et al. Roformer: enhanced transformer with rotary position embedding [J]. Neurocomputing, 2024, 568: 127063
[10] PENG B, QUESNELLE J, FAN H, et al. YaRN: efficient context window extension of large language models [EB/OL]. (2023-09-15)[2025-04-05]. https://arxiv.org/abs/2309.00071
[11] BAI S, CHEN K, LIU X, et al. Qwen2.5-Vl technical report [EB/OL]. (2025-02-19)[2025-04-05]. https://arxiv.org/abs/2502.13923
[12] BLAKEMAN A, BASANT A, KHATTAR A, et al. Nemotron-H: a family of accurate and efficient hybrid mamba-transformer models [EB/OL]. (2025-04-04) [2025-04-05]. https://arxiv. org/abs/2504.03624
[13] GitHub. LLM Hunyuan [EB/OL]. [2025-04-05]. https://github.com/Tencent/llm.hunyuan.T1
[14] WEI J, WANG X, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models [J]. Advances in neural information processing systems, 2022, 35: 24824-24837
[15] LI W, LI Y. Process reward model with q-value rankings [EB/OL]. (2024-10-15)[2025-04-05]. https://arxiv.org/abs/2410.11287
[16] JAECH A, KALAI A, LERER A, et al. Openai o1 system card [EB/OL]. (2024-12-21) [2025-04-05]. https://arxiv. org/abs/2412.16720
[17] SHAO Z, WANG P, ZHU Q, et al. Deepseekmath: pushing the limits of mathematical reasoning in open language models [EB/OL]. (2024-02-05) [2025-04-05]. https://arxiv. org/abs/2402.03300
[18] GUO D, YANG D, ZHANG H, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning [EB/OL]. (2024-01-11)[2025-04-05]. https://arxiv.org/abs/2501.12948
[19] YU Q, ZHANG Z, ZHU R, et al. DAPO: an open-source LLM reinforcement learning system at scale [EB/OL]. (2024-01-11) [2025-04-05]. https://arxiv.org/abs/2503.14476
[20] YUAN Y, YU Q, ZUO X, et al. VAPO: efficient and reliable reinforcement learning for advanced reasoning tasks [EB/OL]. (2025-04-16)[2025-04-25]. https://arxiv.org/abs/2504.05118
[21] ALAYRAC J B, DONAHUE J, LUC P, et al. Flamingo: a visual language model for few-shot learning [J]. Advances in neural information processing systems, 2022, 35: 23716-23736
[22] WU Z, CHEN X, PAN Z, et al. DeepSeek-vl2: mixture-of-experts vision-language models for advanced multimodal understanding [EB/OL]. (2024-12-05) [2025-04-05]. https://arxiv. org/abs/2412.10302
[23] Chameleon Team. Chameleon: mixed-modal early-fusion foundation models [EB/OL]. (2024-01-11) [2025-04-05]. https://arxiv.org/abs/2405.09818
[24] WANG X, ZHANG X, LUO Z, et al. Emu3: next-token prediction is all you need [EB/OL]. (2024-09-23) [2025-04-05]. https://arxiv.

org/abs/2409.18869
[25] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-resolution image synthesis with latent diffusion models [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. IEEE, 2022: 10684-10695
[26] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: transformers for image recognition at scale [EB/OL]. (2020-10-22) [2025-04-05]. https://arxiv. org/abs/2010.11929
[27] LI J, LI D, SAVARESE S, et al. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models [C]//International conference on machine learning. PMLR, 2023: 19730-19742
[28] PEEBLES W, XIE S. Scalable diffusion models with transformers [C]//Proceedings of the IEEE/CVF international conference on computer vision. IEEE, 2023: 4195-4205
[29] WANG X, WANG Z, GAO X, et al. Searching for best practices in retrieval-augmented generation [EB/OL]. (2024-07-11) [2025-04-05]. https://arxiv.org/abs/2407.01219
[30] HUANG Y, CHENG Y, BAPNA A, et al. Gpipe: efficient training of giant neural networks using pipeline parallelism [EB/OL]. (2018-11-16)[2025-04-05]. https://arxiv.org/abs/1811.06965
[31] QI P, WAN X, HUANG G, et al. Zero bubble pipeline parallelism [EB/OL]. (2024-01-11) [2025-04-05]. https://arxiv. org/abs/2401.10241
[32] JACOBS S A, TANAKA M, ZHANG C, et al. Deepspeed ulysses: system optimizations for enabling training of extreme long sequence transformer models [EB/OL]. (2023-09-15) [2025-04-05]. https://arxiv.org/abs/2309.14509
[33] LIU H, ZAHARIA M, ABBEEL P. Ring attention with blockwise transformers for near-infinite context [EB/OL]. (2023-10-11) [2025-04-05]. https://arxiv.org/abs/2310.01889
[34] GALE T, NARAYANAN D, YOUNG C, et al. Megablocks: efficient sparse training with mixture-of-experts [J]. Proceedings of machine learning and systems, 2023, 5: 288-304
[35] MIAO X, WANG Y, JIANG Y, et al. Galvatron: efficient transformer training over multiple gpus using automatic parallelism [EB/OL]. (2022-11-11)[2025-04-05]. https://arxiv.org/abs/2211.13878
[36] MICIKEVICIUS P, STOSIC D, BURGESS N, et al. FP8 formats for deep learning [EB/OL]. (2024-01-11) [2025-04-05]. https://arxiv.org/abs/2209.05433
[37] CHEN C, LI X, ZHU Q, et al. Centauri: enabling efficient scheduling for communication-computation overlap in large model training via communication partitioning [C]//Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems. ACM, 2024: 178-191
[38] LIU A, FENG B, XUE B, et al. DeepSeek-V3 technical report [EB/OL]. (2024-01-11) [2025-04-05]. https://arxiv. org/abs/2412.19437
[39] GitHub. DeepSeek AI: 3FS [EB/OL]. [2025-04-05]. https://github.com/deepseek-ai/3FS
[40] WANG Q, SANG B, ZHANG H, et al. DLRover: an elastic deep training extension with auto job resource recommendation [EB/OL]. (2023-04-04) [2025-04-05]. https://arxiv. org/abs/2304.01468
[41] GLOECKLE F, IDRISSI B Y, ROZIERE B, et al. Better & faster large language models via multi-token prediction [EB/OL]. (2024-04-30)[2025-04-05]. https://arxiv.org/abs/2404.19737
[42] LU E, JIANG Z, LIU J, et al. MoBA: mixture of block attention for long-context LLMs [EB/OL]. (2024-01-11)[2025-04-05]. https://arxiv.org/abs/2502.13189
[43] CHEN Q, QIN L, LIU J, et al. Towards reasoning era: a survey of long chain-of-thought for reasoning large language models [EB/OL]. (2024-01-11) [2025-04-05]. https://arxiv. org/abs/2503.09567
[44] 韩银俊, 牛家浩, 屠要峰. 数据管理系统发展趋势与挑战 [J]. 中兴通讯技术, 2023, 27(2): 64-71. DOI: 10.12142/ZTETJ.202304012
[45] 田海东, 张明政, 常锐, 等. 大模型训练技术综述 [J]. 中兴通讯技术, 2024, 30(2): 21-28. DOI: 10.12142/ZTETJ.202402004
[46] 韩炳涛, 刘涛. 大模型关键技术与应用 [J]. 中兴通讯技术, 2024, 30(2): 76-88. DOI: 10.12142/ZTETJ.202402012
[47] FENG B Y, FENG M X, WANG M R, et al. Multi-agent hierarchical graph attention reinforcement learning for grid-aware energy management [J]. ZTE Communications, 2023, 21(3): 11 - 21. DOI: 10.12142/ZTECOM.202303003
[48] GitHub. Model context protocol [EB/OL]. [2025-04-05]. https://github.com/modelcontextprotocol
[49] GitHub. Agent2Agent (A2A) protocol [EB/OL]. [2025-04-05]. https://github.com/google/A2A
[50] 邵宏, 谢大雄. 具身智能机器人技术 [J]. 中兴通讯技术, 2024, 30(S1): 40-44. DOI: 10.12142/ZTETJ.2024S1006
[51] KIM M J, PERTSCH K, KARAMCHETI S, et al. Openvla: an open-source vision-language-action model [EB/OL]. (2024-01-11) [2025-04-05]. https://arxiv.org/abs/2406.09246
[52] YANG Z, LI L, LIN K, et al. The dawn of LMMs: preliminary explorations with GPT-4V (ision) [EB/OL]. (2023-09-29) [2025-04-05]. https://arxiv.org/abs/2309.17421
[53] Dify. Build production: ready agentic AI solutions [EB/OL]. [2025-04-05]. https://dify.ai/
[54] RAGFlow. Build generative AI into your business [EB/OL]. [2025-04-05]. https://ragflow.io/

作 者 简 介

包义明，中兴通讯股份有限公司人工智能算法工程师、中兴通讯“蓝剑计划”员工；主要研究方向为计算机视觉、AI Infrastructure、多模态大模型等。

林阳，中兴通讯股份有限公司技术预研工程师；主要研究方向为人工智能、数据库、大模型技术及应用等。

屠要峰，中兴通讯股份有限公司中心研究院副院长、中国人工智能学会理事、中国计算机学会杰出会员、大数据/数据库/信息存储专委会常务委员；主要研究方向为大数据、人工智能、数据库及存储。