大模型作为近几年的爆点方向，引起各大厂商、学术界的极大关注。作为一个发展前景广阔的方向，有必要去了解该方向的发展。本文主要汇总了大模型推理的相关技术、业界/学术界以及我司内部一些前辈整理的相关材料，希望对各位的工作能够有所帮助。

**自由讨论详情**

* [0 引子](https://jx.huawei.com/community/#0-%E5%BC%95%E5%AD%90)
* [1 大模型框架介绍](https://jx.huawei.com/community/#1-%E5%A4%A7%E6%A8%A1%E5%9E%8B%E6%A1%86%E6%9E%B6%E4%BB%8B%E7%BB%8D)
* [2 LLM Serving](https://jx.huawei.com/community/#2-llm-serving)
* [3 大模型内存管理](https://jx.huawei.com/community/#3-%E5%A4%A7%E6%A8%A1%E5%9E%8B%E5%86%85%E5%AD%98%E7%AE%A1%E7%90%86)
* [4 大模型批处理与调度](https://jx.huawei.com/community/#4-%E5%A4%A7%E6%A8%A1%E5%9E%8B%E6%89%B9%E5%A4%84%E7%90%86%E4%B8%8E%E8%B0%83%E5%BA%A6)
* [5 PD分离](https://jx.huawei.com/community/#5-pd%E5%88%86%E7%A6%BB)
* [6 大模型存储](https://jx.huawei.com/community/#6-%E5%A4%A7%E6%A8%A1%E5%9E%8B%E5%AD%98%E5%82%A8)
* [7 大模型与推荐系统](https://jx.huawei.com/community/#7-%E5%A4%A7%E6%A8%A1%E5%9E%8B%E4%B8%8E%E6%8E%A8%E8%8D%90%E7%B3%BB%E7%BB%9F)
* [8 ARM & LLM推理](https://jx.huawei.com/community/#8-arm--llm%E6%8E%A8%E7%90%86)
* [9 昇腾大模型解决方案](https://jx.huawei.com/community/#9-%E6%98%87%E8%85%BE%E5%A4%A7%E6%A8%A1%E5%9E%8B%E8%A7%A3%E5%86%B3%E6%96%B9%E6%A1%88)
* [10 端侧推理](https://jx.huawei.com/community/#10-%E7%AB%AF%E4%BE%A7%E6%8E%A8%E7%90%86)

**个人收集的资料难免有所欠缺，欢迎各位看官补充、完善，真正实现新手速通大模型。**

1. **0 引子**
2. **1 大模型框架介绍**

[1] 复旦大学-大模型从理论到实践书籍:<https://intro-llm.github.io/chapter/LLM-TAP.pdf>  
[2] 中国人民大学-大语言模型:<https://github.com/LLMBook-zh/LLMBook-zh.github.io/blob/main/LLMBook.pdf>

1. **2 LLM Serving**

[1] <https://www.youtube.com/watch?v=TJ5K1CO9Wbs>  
[2] LLM Inference Serving: Survey of Recent  
Advances and Opportunities: <https://arxiv.org/pdf/2407.12391>  
[3] Awesome-LLM-System-Papers:<https://github.com/AmadeusChan/Awesome-LLM-System-Papers>  
[4] 各类大模型对比及相关应用分享：[https://3ms.huawei.com/km/groups/3956141/blogs/details/15122976](https://3ms.huawei.com/km/groups/3956141/blogs/details/15122976" \t "_blank)  
[5] LLM 的推理优化技术纵览:<https://zhuanlan.zhihu.com/p/642412124>  
[6] A Survey on Efficient Inference for Large Language Models: [https://arxiv.org/pdf/2404.14294;https://finance.sina.cn/tech/2024-06-13/detail-inayqwqe9277049.d.html?fromtech=1&from=wap](https://arxiv.org/pdf/2404.14294;https:/finance.sina.cn/tech/2024-06-13/detail-inayqwqe9277049.d.html?fromtech=1&from=wap)  
[7]昇腾大模型关键特性（推理篇）- 推理框架＆MindIE：[https://jx.huawei.com/community/comgroup/postsDetails?postId=748a9baa837b41108447e024dfb3531a&noTop=true&type=freePost](https://jx.huawei.com/community/comgroup/postsDetails?postId=748a9baa837b41108447e024dfb3531a&noTop=true&type=freePost" \t "_blank)  
[8] 业界大模型推理框架的分析——以NVIDIA和我司为例:<https://jx.huawei.com/community/comgroup/postsDetails?postId=236af386d51243a99232fda49fa9ea00>  
[9] 大模型加速推理技术总结和展望:<https://jx.huawei.com/community/comgroup/postsDetails?postId=6cf7d5fcf1a14cbf9338dc352933bbdd>  
[10]AI模型推理加速前沿技术讨论合集:<https://jx.huawei.com/community/comgroup/postsDetails?postId=f00bcf1d0384434093b23fcfd408ae8e>  
[11] Mastering LLM Techniques: Inference Optimization: <https://developer.nvidia.com/blog/mastering-llm-techniques-inference-optimization/>  
[12]王均松：搜推和大模型推理场景对内存、通信及其池化的需求:<https://jx.huawei.com/community/comgroup/postsDetails?postId=9e7a5a34cd0e4da299f2e798005c5208>

LLM Serving基本架构

LLM Serving主要关注点

LLM Serving已有系统对比与主流技术趋势

1. **3 大模型内存管理**

[0] 图解大模型计算加速系列之：vLLM核心技术PagedAttention原理:<https://mp.weixin.qq.com/s?__biz=Mzg2NjcwNjcxNQ==&mid=2247485614&idx=1&sn=5600ea665d942b7ff290caded1e2252f&chksm=ce47fcdaf93075cc4582bfb15eb822840c332d56122df5b59bd8722426d07ac5b097cf032ba4&scene=21#wechat_redirect>  
[1] Li S, Xue F, Baranwal C, et al. Sequence parallelism: Long sequence training from system perspective[J]. arXiv preprint arXiv:2105.13120, 2021. arXiv:2309.14509, 2023. <https://arxiv.org/abs/2105.13120>  
[2] Sheng Y, Zheng L, Yuan B, et al. Flexgen: High-throughput generative inference of large language models with a single gpu[C]//International Conference on Machine Learning. PMLR, 2023: 31094-31116. <https://arxiv.org/abs/2303.06865>  
[3] Lin B, Peng T, Zhang C, et al. Infinite-LLM: Efficient LLM Service for Long Context with DistAttention and Distributed KVCache[J]. arXiv preprint arXiv:2401.02669, 2024. <https://arxiv.org/abs/2401.02669>  
[4] MiniCache 和 PyramidInfer 等 6 种优化 LLM KV Cache 的最新工作：[https://www.51cto.com/aigc/913.html](https://www.51cto.com/aigc/913.html" \t "_blank)  
[5] LLM时代中的AI推理优化：[https://blog.csdn.net/jinzhuojun/article/details/139693922](https://blog.csdn.net/jinzhuojun/article/details/139693922" \t "_blank)  
[6] LLM推理KVcache架构洞察:<https://jx.huawei.com/TI/report/details/1e5addedeeab4b218252d1dd63af2a51>  
[7] 长序列场景下大模型推理中KVCache机制还可以怎么优化:<https://jx.huawei.com/community/comgroup/postsDetails?postId=0c870eece7054a4593fafe1a2edec7ff>  
[8] 面向AI推理加速的分离式内存技术: <https://jx.huawei.com/community/comgroup/postsDetails?postId=fa69f7fd54944517a30bdca0f383603e>

1. **4 大模型批处理与调度**

[1] <https://towriting.com/blog/2023/09/02/continuous-batch/>  
[2] 大模型continuous Batching详解：[https://blog.csdn.net/qq\_27590277/article/details/135710435](https://blog.csdn.net/qq_27590277/article/details/135710435" \t "_blank)  
[3] LLM大语言模型推理引擎/系统级优化 - 调度、并行、KVCache、MOE优化理论与方法:<https://jx.huawei.com/TI/report/details/d523bc598f4446a1ac32f77f1a33131e>

1. **5 PD分离**

[1] 大模型云化推理之PD分离部署：[https://jx.huawei.com/community/comgroup/postsDetails?postId=e64c6629378446208c7fd166a545f787&noTop=true&type=freePost&zoneId=5&target=zone\_all](https://jx.huawei.com/community/comgroup/postsDetails?postId=e64c6629378446208c7fd166a545f787&noTop=true&type=freePost&zoneId=5&target=zone_all" \t "_blank)  
[2] TetriServe: 像叠俄罗斯方块一样优化大模型推理, 提升性价比：[https://jx.huawei.com/community/comgroup/postsDetails?postId=d21e9ca7dc6e4a3486c47b4fb318fb79&noTop=true&type=freepost](https://jx.huawei.com/community/comgroup/postsDetails?postId=d21e9ca7dc6e4a3486c47b4fb318fb79&noTop=true&type=freepost" \t "_blank)  
[3] 大模型PD分离推理收益分析和模拟验证：[https://jx.huawei.com/community/comgroup/postsDetails?postId=9bf3a5223fef4881b5851b94e9c013c4](https://jx.huawei.com/community/comgroup/postsDetails?postId=9bf3a5223fef4881b5851b94e9c013c4" \t "_blank)  
[4] 大模型PD分离推理系列论文分析：[https://jx.huawei.com/community/comgroup/postsDetails?postId=48d26d0d97ee42c8be9f6707d58d27d1](https://jx.huawei.com/community/comgroup/postsDetails?postId=48d26d0d97ee42c8be9f6707d58d27d1" \t "_blank)  
[5] MindIE PD分离实践：[https://jx.huawei.com/community/comgroup/postsDetails?postId=18709d02425d4c429ce604a0651ef624&noTop=true](https://jx.huawei.com/community/comgroup/postsDetails?postId=18709d02425d4c429ce604a0651ef624&noTop=true" \t "_blank)

1. **6 大模型存储**

[1] AI大模型场景下的存储系统技术演进趋势（清华-舒继武）:<https://jx.huawei.com/community/comgroup/postsDetails?postId=c276c141f1ae4287802105f4afb7dff8&noTop=true&type=freePost&zoneId=3&target=zone_all>  
[2] 人工智能训推一体系统架构（上交-过敏意）:<https://jx.huawei.com/community/comgroup/postsDetails?postId=1cb3b6ae148b466480ff44b5d4c038b1&noTop=true&type=freePost&zoneId=3&target=zone_all>

1. **7 大模型与推荐系统**

[0] 全链路解析！推荐系统技术综述:[https://blog.csdn.net/Kaiyuan\_sjtu/article/details/130758200；](https://blog.csdn.net/Kaiyuan_sjtu/article/details/130758200%EF%BC%9B)  
23张图，带你入门推进系统：[https://www.woshipm.com/pd/4223123.html](https://www.woshipm.com/pd/4223123.html" \t "_blank)  
[1] 华为：大语言模型在推荐系统的实践应用：[https://zhuanlan.zhihu.com/p/674032466](https://zhuanlan.zhihu.com/p/674032466" \t "_blank)  
[2] CNCC 2023 | 华为唐睿明：推荐系统如何从大语言模型中取长补短：从应用视角出发:<https://www.bilibili.com/video/BV1YN4y167Vy/?vd_source=9d824af45643ac7c36dece6c6df7e237>  
[3] 预训练大模型下的搜索推荐系统范式：[https://3ms.huawei.com/km/groups/2027333/blogs/details/14841243](https://3ms.huawei.com/km/groups/2027333/blogs/details/14841243" \t "_blank)  
[4] LLM在电商推荐系统的探索与实践:<https://mp.weixin.qq.com/s/QFV1kJ6ElGyvHK4mXLJNxg>  
[5] LLM for Recommendation Systems:<https://github.com/WLiK/LLM4Rec-Awesome-Papers>  
[6] 搜索推荐广告技术洞察：[https://jx.huawei.com/TI/report/details/60cf3550b2d64e8096f266cf6c4369e5](https://jx.huawei.com/TI/report/details/60cf3550b2d64e8096f266cf6c4369e5" \t "_blank)

1. **8 ARM & LLM推理**

[1] Full-Stack Optimizing Transformer Inference on ARM Many-Core CPU: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10138027&tag=1>  
[2] CPU跑AI，不被时代抛下的自救:<https://jx.huawei.com/community/comgroup/postsDetails?postId=fa1366893a664b3bb1c94a1dfcaeb6a5>  
[3] 端侧模型推理框架调研:<https://3ms.huawei.com/km/groups/3952228/blogs/details/12685307>  
[4] 1个1400行cpp，1行编译命令，让每台鲲鹏920能几乎满带宽地推理LLaMa系列大模型(demo,代码)：<https://jx.huawei.com/community/comgroup/postsDetails?postId=c7b399c67ec9468293ee6b34f64e8285>

1. **9 昇腾大模型解决方案**

[1] 昇腾大模型解决方案主打胶片：<https://3ms.huawei.com/documents/docinfo/906078184359968768>  
[2] 昇腾AI推理解决方案技术主打胶片(视图推理):<https://3ms.huawei.com/documents/docinfo/906078184359968768>

1. **10 端侧推理**

[1] PowerInfer-2: Fast Large Language Model Inference on a Smartphone: <https://powerinfer.ai/v2/>