

# 大模型推理介绍



KV-cache transfer

. . .

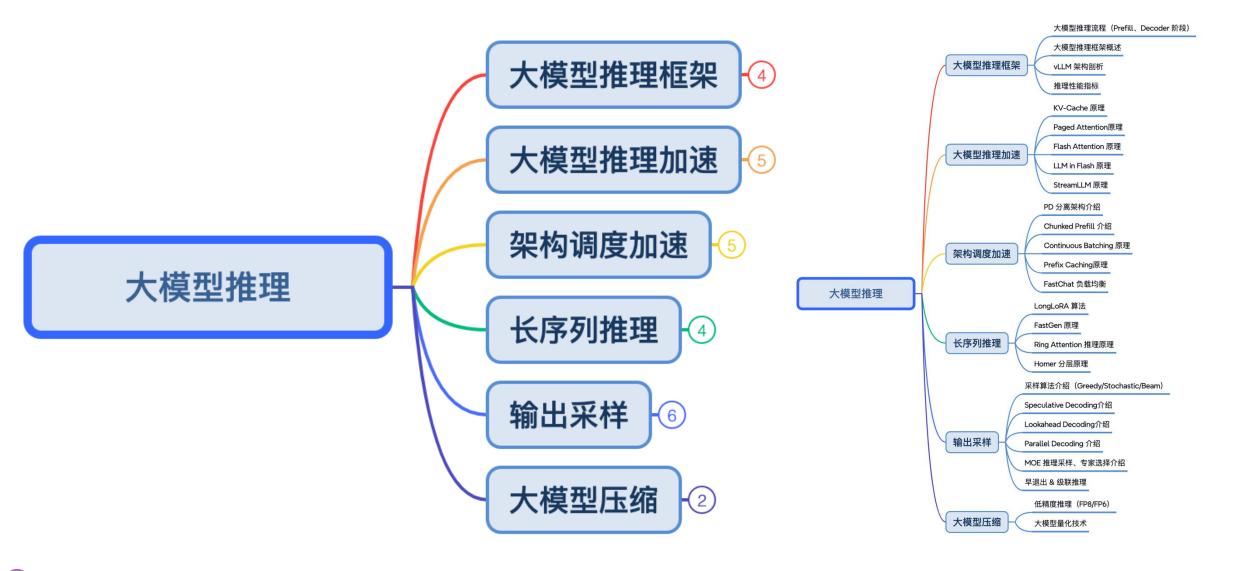
## 思考

- 大模型推理框架跟传统推理框架,有什么不同?
  - · 大模型推理框架只推理大模型,聚焦 Transformer 架构
  - 。 大模型推理要包含服务化, 主要在云端使用
  - 。 ??? 还有吗? 那么架构上有什么区别?





# 目录





大模型推理框架

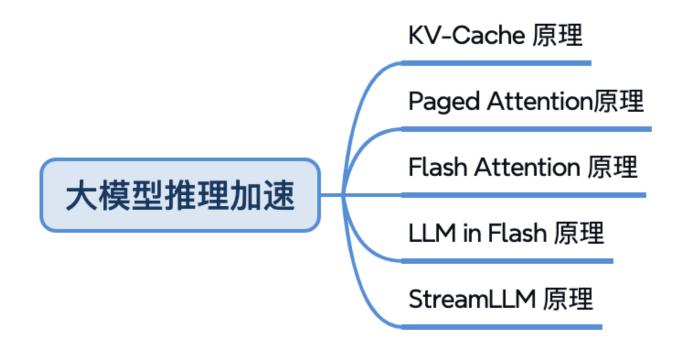
大模型推理流程 (Prefill、Decoder 阶段)

大模型推理框架概述

vLLM 架构剖析

推理性能指标







架构调度加速

PD 分离架构介绍

Chunked Prefill 介绍

Continuous Batching 原理

Prefix Caching原理

FastChat 负载均衡



LongLoRA 算法
FastGen 原理
K序列推理
Ring Attention 推理原理
Homer 分层原理



输出采样

采样算法介绍(Greedy/Stochastic/Beam)

Speculative Decoding介绍

Lookahead Decoding介绍

Parallel Decoding 介绍

MOE 推理采样、专家选择介绍

早退出 & 级联推理

低精度推理 (FP8/FP6)

大模型压缩

大模型量化技术





把AI系统带入每个开发者、每个家庭、 每个组织,构建万物互联的智能世界

Bring Al System to every person, home and organization for a fully connected, intelligent world.

Copyright © 2024 XXX Technologies Co., Ltd. All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. XXX may change the information at any time without notice.



 $Git Hub\ https://github.com/chenzomi I\ 2/AIF oundation$