一种私有化大语言模型生态服务的高效部署方法及系统

Principal Investigator

Mr J X Wang Software Engineer Manager, AIFT

Registered Patent Region	Patent/ Filing No.	Description	Issued/ Filling Date	Status
China	202411584177.0	一种私有化大语言模型生态服务的高效部署方法及系统	7 Nov 2024	Under Examination
Hong Kong	HK30118138	一种私有化大语言模型生态服务的高效部署方法及系统	6 Jun 2025	Granted

Stage of Technology Transfer: Commercially Viable Technology/ Product Ready for Sale
Research Area: Large Language Model, Multimodal Data Processing, Cluster Management, Dynamic Load Balancing, and Performance Monitoring

Abstract

This invention provides an efficient deployment method and system for a privatized large language model (LLM) ecosystem service. The system includes an evaluation module, metrics collector, monitoring module, vLLM framework, model cluster service module, and request distribution module. The evaluation module performs quantitative assessments of system performance on different hardware platforms. The metrics collector gathers real-time performance metrics, resource status, and load information from server clusters. The monitoring module oversees the server cluster operations. The vLLM framework dynamically allocates computational resources, managing multi-modal server clusters and utilizing PM2 for unified scheduling to enable efficient parallel processing of multiple model tasks. The model cluster service module launches the vLLM framework, encapsulating various LLM instances within Docker containers. The request distribution module adapts request allocation in real time based on the cluster's resource status and load information, as provided by the metrics collector. This system achieves intelligent request distribution and dynamic load balancing, facilitating integration and efficient management of multi-modal large language models.

摘要

本发明涉及一种私有化大语言模型生态服务的高效部署方法及系统，包括评估模块、指标采集器、监控模块、vLLM 框架、模型集群服务模块和请求分发模块。评估模块对系统在不同硬件平台下的性能进行量化评估；指标采集器实时采集服务器集群性能指标、集群资源状态与负载信息；监控模块对服务器集群进行监控；vLLM 框架动态分配计算资源，实现对多模态服务器集群的管理，并通过 PM2 进行统一管理和调度实现多模型高效并行处理任务；模型集群服务模块启动 vLLM 框架，将不同的大模型实例封装在 Docker 容器中；请求分发模块根据指标采集器实时反馈的集群资源状态与负载信息自适应地分发外部请求。实现了智能化请求分发与动态负载均衡，多模态大语言模型的集成与高效管理。