文件编号		SSAD-02-01	
版本号	1.0	创建日期	2024-06-01
作 者	Haixia Pan Huobin Tan	更新日期	2024-09-01

“大模型集成训练平台”背景陈述

文档说明：本案例根据作者所参与的项目经历和部分网络资料虚构而成，专门用于“软件系统分析与设计”课程案例教学，案例中可能有很多细节考虑不周，欢迎读者针对案例中的问题与作者沟通和探讨（haixiapan@buaa.edu.cn，thbin@buaa.edu.cn）；另有部分细节没有阐述的，读者可结合自己的理解酌情考虑。未经作者许可，请勿随意在网络上传播和转载，更不能用于其他商业用途。

AI 技术的迅速发展，如何有效地实现大规模模型训练与部署，并进一步推动人工智能技术的普及和应用，已经是科学技术领域的热门研究领域。

大模型集成训练平台是专门为企业和开发者设计的，用于快速构建、部署和应用大规模人工智能模型的服务。这些平台通常提供一站式服务，包括模型训练、优化、在线服务部署等功能，并支持多种开源大模型的接入和适配。平台的目标是通过简化操作流程和提供灵活的资源配置，帮助用户更高效地进行大模型训练和推理。

本项目定位建设一套面向企业和开发者设计的大模型集成训练平台。该平台由专门的管理机构运行和维护，按照用户需求和技术标准进行配置管理。其基本的运行和管理机制包括以下几个方面：

1. 多模型集成与支持：平台应集成多种神经网络的训练和推理功能，包括但不限于大语言模型和视觉模型，如 LLaMA、Qwen、YOLO 等。用户通过 Web 界面配置参数，选择所需的模型及训练环境，启动模型训练或推理任务，可参考[1][2][3]。例如，用户在 web 页面，选择基础模型、选择数据集、选择显卡编号、设置 batch size，epoch 等参数，即可启动模型训练。

2. 资源调度与管理：普通用户可以在多卡环境中选择显卡资源，启动模型训练。系统管理员负责配置普通用户可使用的显卡编号，并指定其可以启动的任务类型。例如，管理员可以限制用户 A 仅使用显卡 2 和显卡 3，并只能启动大语言模型的微调任务，而无法启动视觉模型的训练。

3. 任务监控与日志管理：平台需提供实时的任务监控和日志记录功能。用户可以通过 Web 界面实时查看训练进度、loss 变化等关键指标。同时，平台应支持日志导出和分析功能，以便用户在训练任务完成后进行详细的数据分析和调优。

4. 数据管理与存储：因为多源异构数据需要定制化规则较多，所以本平台可以支持数据预处理功能。平台应支持数据集的集中管理和高效存储。用户也可以上传已经改好格式的数据集，即无需处理，直接用于模型训练。

5. 扩展与升级：系统一期定位于多种模型的训练、推理等基本需求，后期成熟后，可以考虑提供有偿增值服务，如付费使用高性能计算资源、专业的模型优化咨询服务等，以进一步增强平台的商业价值。

通过上述功能和机制的实现，大模型集成训练平台将能够为用户提供一个高效、灵活、安全的训练和推理环境，助力人工智能技术的快速发展与应用。

[1] https://github.com/hiyouga/LLaMA-Factory，该仓库实现了 LLM 的训练、推理功能，ui 界面较为简单
[2] 阿里 PAI 平台：https://help.aliyun.com/zh/pai/
[3] 百度千帆平台：https://console.bce.baidu.com/qianfan/overview