

项目名称：面向人工智能模型的数据治理软硬件协同关键技术研究

一、背景（50字以内）

智能涌现归结于大数据和大知识建设，数据治理是人工智能发展的基石，需要满足高质量、海量性、智能化的需求。

二、拟解决的科学/技术问题（50字以内）

基于异构计算平台对海量数据进行高性能智能优化、治理，构建全面、高质量的数据体系，显著提升人工智能模型表现。

三、研究内容（80-200字，针对……研究……）

针对人工智能模型对海量数据的需求，构建高性能高端服务器异构计算平台；研究基于影响力函数的多源异构数据整合方法，构建全面的数据体系，成倍扩充数据的广度和深度；研究基于深度学习模型的智能化数据质量检查和精准的数据调优；研究高效的交互式数据治理 workflow 生成，提出基于强化学习的智能化用户 workflow 推荐方法；研究端对端的数据治理 workflow 优化方法，通过数据物化与重用、多设备交互执行加速等技术，提升数据治理 workflow 的效率。

四、考核指标（80-200字，实现……提出……交付……发表……）

实现一套支持大规模数据体系建设与调优方案的高算力硬件系统架构设计，支持异构缓存协议 CXL 内存扩展技术和文件系统透明压缩异构加速技术，整机搭载算力不低于

600 TOPS；针对 ChatGLM 和 Stable Diffusion 大模型构建预训练和微调数据体系，数据量不少于 8 T；通过数据治理实现模型在 MMLU 和 C-Eval 基准上准确率平均提升 10%以上；提出交互式数据治理 workflow 生成与智能推荐方案，通过离散 CPU、GPU 混合架构细粒度交互运行等技术，使数据治理 workflow 在相同单机实验环境下的执行效率提升 10 倍以上；发表高水平学术论文不少于 4 篇，申请专利不少于 6 项。

五、经费需求及优势团队

经费需求：120 万

优势团队：