2024 年度朝阳区人工智能关键技术研发课题 征集方向

一 、以运筹学为基础的国产核心算法商用软件研发与应用

**（ 一 ）研究内容**

面向北京市能源电力、智能制造、轨道交通等重点行业的协同调 度、需求预测、运营决策以及数据安全等方面的实际需求 ，开展以运 筹学为基础的国产化可商用核心算法软件开发攻关 ，研究针对复杂业 务场景的国产化底层计算求解器 ；研究面向重点行业生产经营管理、 运营销售管理、库存优化管理等关键节点的国产化人工智能软件 ；研 究以机器学习和运筹优化算法为基础的智能辅助决策软件；面向能源、 制造、交通等重点行业，解决典型应用场景需求，打造标杆应用场景。

**（ 二 ） 考核指标**

1. 开发基于运筹学的国产智能计算求解器软件，求解器应具备 原生变成语言接 口、第三方建模语言接 口 ，支持读写主流的标准格式 模型文件 ，支持优化类型包括线性规划和混合整数规划问题。求解器 应支持亿量级求解， 求解速度与国外求解器对比提升10%。

2. 项目研究成果与不少于 3 家行业头部企业签订项目应用实施 的服务合同。

3. 申请不少于 1 个专利 ； 申请不少于 3 个软件著作权。

二 、在消费级显卡上进行大模型训练和推理的软件平台研发与应用

**（ 一 ）研究内容**

研发在消费级显卡上进行大模型训练和推理的软件平台 ，研究内 利用消费级 GPU 对大模型进行预训练、微调和强化学习的方法 ，包括 对显存占用的持续优化，对注意力机制及 CUDA 算子深入调优，以及自 动化数据选择方法优化等研究内容， 目标实现用更少的显存、更快的 速度、更少的数据实现更优的模型训练效果；研究大模型在消费级 GPU 上提升推理吞吐量的方法 ，包括并发策略、MoE 结构、专家路由设计 等研究内容 ；研究大模型在消费级 GPU 上优化通信显存负载的方法 ， 包括对通信进行多级分层优化 ，将通信过程与计算进行重叠 ，减少通 信等待时间等研究内容。

**（ 二 ）考核指标**

1. 支持 GLM/Qwen/Baichuan/InternLM/Mistral/LlaMa 模型 ，支持 DeepSeek/Qwen/Mixtral/DataBricks 等多种 MoE 模型架构。

2. 实现在 8 卡 4090 及同等性能配置服务器上， 完成千亿参数 级别模型的预训练、推理、强化学习过程。

3. 实现千亿参数模型在 8 卡 4090 及同等性能配置服务器上， 推理速度>3000 token/s。

4. 支持多种模型量化方式 ，量化后评测效果差异<1%。

5. 申请不少于 1 个专利 ； 申请不少于 3 个软件著作权。

三 、解决大规模数据集分析问题的基础决策模型研发与应用

**（ 一 ）研究内容**

研究面向企业发展所需要的大规模商家特征分析需求 ，研发基础 决策模型。研究国产 GPU 配套算子库、数学库、运算库 ，提升智能决 策算法工具应用自主可控能力： 算子库方面 ，通过 TVM 和 OneFlow 等 工具 ，开发针对国产GPU 优化的数学运算和逻辑运算函数 ，提升深度 学习框架的效率 ；数学库方面 ，开发高性能数学库 ，支持科学计算和 工程计算中的矩阵运算和优化算法 ；运算库方面 ，借鉴 NVIDIA cuDNN 和 OpenCL 框架，开发高性能运算库，提高深度学习模型训练和推理效 率。基于上述研究内容 ，在本地生活产业的复杂业务场景中 ，研究规 模数据智能建模的方法 ，开发基础决策模型。

**（ 二 ）考核指标**

1． 开发基础决策模型 ，包含大数据处理模块、决策算法模块以 及业务场景解决方案模块。

2． 模型可支持百万级元数据处理、千万级数据计算规模。

3． 模型精确率 Precision 达 10%、召回率 Recall 达 80%、准确率 Accuracy 达 40%、 区分度 AUC 达 0.8。

四 、大模型驱动的智能风控管理平台研发与应用

**（ 一 ）研究内容**

在金融领域 ，研发大模型驱动的智能风控管理平台 ，研究通过多 源异构数据（如交易数据、社交网络数据、行为数据等）的高效整合， 构建高维特征空间 ，提升风控模型的泛化能力和准确性 ；研究基于图 神经网络（GNN）的关系网络分析技术，开发识别潜在风险和异常模式 的算法，提升信用评分模型的准确性；研究构建自适应风险控制系统， 动态调整风控策略， 实现实时数据监控与异常检测。

**（ 二 ）考核指标**

1.平台需支持自定义决策流设计， 允许用户灵活组合多个风险模 型， 以适应不同的业务场景和需求 。平台应具备模型管理和动态更新 能力，支持不少于10 种模型的集成应用。决策流的配置和部署时间不 超过 5 分钟 ，支持 A/B 测试和实时策略优化。

2.支持从海量数据中自动提取风险相关标签 ，标签提取准确率≥ 90% ，并实现自动化分类和风险预警功能。

3.基于分布式计算的模型训练时间减少30%以上，支持千万级用户 规模的数据处理 ，每个用户拥有万维特征的复杂特征集。

4. 申请不少于 2 个软件著作权。