# 服务内容

**AI计算优化服务**针对客户未适配昇腾平台的模型，提供面向昇腾平台的迁移调优服务。本服务旨在帮助客户将自有模型快速落地昇腾AI平台，帮助客户评估模型迁移至昇腾AI平台的可行性，基于具体模型算法与框架特点，制定并实施模型迁移方案，针对昇腾模型进行计算精度调优与计算性能调优。

## 模型迁移

提供专属服务工程师，将客户模型从三方AI平台迁移至昇腾AI平台，具体包含以下内容：

* 模型信息采集

调研客户业务需求与场景，分析客户的目标模型，明确客户模型运行的硬件平台的类型及具体型号，获取待优化的模型源码、参数配置、模型权重测试数据集等文件。

获取与分析模型优化前的硬件平台信息与迁移前的模型计算精度与性能参考基线。

* 模型迁移优化方案设计

制定模型迁移的具体实施方案。结合模型迁移的目标硬件与目标计算框架类型（PyTorch/MindSpore/MindIE/MindSpeed），设计模型加速适配改造方案和具体的优化实施方案。并向客户澄清模型迁移的具体配合事项和验收标准。

* 模型迁移实施

按照模型迁移的整体实施方案进行模型迁移。搭建模型迁移需要的开发环境，安装模型依赖与模型套件，配置模型并行训练需要的网络环境；

模型训练适配包含模型并行训练环境搭建，训练数据加载适配，模型网络搭建，损失函数适配，优化器适配，训练/推理/评估脚本适配等服务内容，将模型从三方平台迁移至昇腾AI平台进行模型预训练或微调。

模型推理加速需根据业务场景，选用离线推理方案、MindIE-LLM方案，如果没有推理加速需求，可选用在线推理方案。离线推理方案含模型推理加速环境搭建，离线模型转换，离线模型前向推理代码适配等服务内容； MindIE-LLM方案包含模型推理环境搭建，模型组图分析，ATB加速库开发，加速库调用与执行等服务内容；在线推理方案包含模型推理环境搭建，模型网络搭建，模型前向推理适配等服务内容。

## 模型调优

针对已迁移模型优化模型的计算精度与计算性能至服务验收标准。

* 模型精度调优

模型迁移适配完成后，加载权重进行模型推理，对齐模型前向计算结果，减少因平台切换产生计算误差。如果在模型训练过程中如果出现精度不达标或收敛效果差的情况，提供模型精度调优服务，主要包含以下内容：

1. 对齐模型前向推理过程中的计算结果。

2）分析模型训练日志、模型网络输出，采集精度数据，定位模型训练过程中的精度异常、数据溢出等问题，保障模型训练损失曲线与三方参考硬件的训练损失曲线平均相对误差小于2%。

3）优化后模型在评测数据集的得分与迁移前的模型推理在评测数据集的得分误差小于2%（基于项目实质、实际测试集情况填写）

* 模型性能调优

模型通过精度优化后，分析模型训练/推理过程中出现的性能瓶颈，优化模型计算性能至合同约定基线，主要包括如下服务内容：

1）性能数据分析与采集：部署性能分析工具，获取框架调度过程与CANN下发执行过程中的性能数据，分析性能瓶颈。

2）模型训练性能优化：有选择性地运用参数配置优化、内存优化、并行策略优化、融合算子优化、调度优化和通信优化等手段提升模型训练的性能。

3）模型推理性能优化：有选择性的运用参数配置优化、分布式推理、Kernel优化、KVCache、内存IO优化、Continuous Batching等特性调用等手段提升模型推理性能。

迁移后模型在通过性能调优后，达到合同约定的要求。

## 模型验收

基于客户提供的昇腾AI平台和验证环境，使用业界开源或客户提供的数据集对迁移后的模型进行测试。在昇腾平台上完成模型训练任务的功能验证，模型参数能正常更新，训练完成后能正常生成权重文件；模型迁移后的精度与性能指标满足与客户达成的目标。验收通过后，输出模型迁移测试报告，提供可直接在昇腾AI平台训练的模型代码、训练脚本与使用说明文档。

# 服务价值

## 模型迁移概述

### 什么是模型迁移

将原本设计用于GPU或其他三方平台的深度学习模型训练代码，经过模型代码修改等适配操作，来适应NPU的架构和编程，让模型能在NPU上进行高性能运行。

### 为什么要做模型迁移

在将模型从其他三方平台迁移到NPU时，由于硬件架构和库的不同，涉及到一系列底层到上层的适配操作。以GPU为例，模型迁移至NPU需要适配的原因可分为三方面：

**硬件特性和性能特点差异：**由于NPU和GPU的硬件特性和性能特点不同，模型在NPU上可能需要进一步的性能调试和优化，以充分发挥NPU的潜力。

**计算架构差异**：CUDA（Compute Unified Device Architecture）+ CuDNN是NVIDIA GPU的并行计算框架，而CANN（Compute Architecture for Neural Networks）是华为NPU的异构计算架构。

**深度学习框架差异**：为了支持NPU硬件，需要对PyTorch框架进行适配：包括适配张量运算、自动微分等功能，以便在NPU上高效执行。PyTorch正在原生支持NPU 。

## 客户收益

### 模型迁移，加速业务上线

华为资深AI专家团队针对模型进行源码级迁移，针对模型训练、微调、评测、推理过程提供计算加速服务，支撑三方模型快速适配昇腾平台。AI计算加速，需要依赖大量前后端专家的持续投入与深度参与，人才梯队构筑难，引入华为原厂服务可以在自有人才构筑阶段期间提前引入模型迁移调优能力，加速业务上线，保障业务高可用。

### 计算精度优化，保障业务连续性

模型迁移时，推理场景会逐算子对比模型计算精度，模型训练过程中对齐损失函数计算结果，保障模型迁移前后计算精度不受计算平台影响，保障业务连续性。

### 专业性能调优，降低算力消耗

从算子、内存、通信、调度等维度调优，调优效率提升50%，平均模型性能调优20%以上；提升模型训练效率可显著降低模型训练成本，以GPT4\_1800B为例，训练成本为7800万美金，每提升1%的MFU（集群利用率），可节省约130万美金。提升模型推理效率，可在同样的设备数量，服务质量下，覆盖更多的用户，提升模型成本竞争力。

