昆仑技术（全称：河南昆仑技术有限公司）成立于 2022 年 10 月，坚持以客户为中心，继承全球领先的技术、市场和服务能力，持续为客户和伙伴创造价值，助力政务、运营商、金融等行业的数字化转型。公司依托自主创新的软硬件研究中心、先进智能制造以及开放的联合创新实验室等核心能力，响应国家战略，聚焦通用计算、AI 异构计算、基础软件与数据服务等领域，提供稳定可靠、安全可信、绿色可持续的 IT 产品与创新解决方案。

方案名称：KunLun Space 大模型加速引擎
提供单位：河南昆仑技术有限公司
方案介绍：

一、背景：

随着人工智能技术的飞速发展，AI 大模型已经成为推动科技进步和行业创新的重要力量。然而，在 AI 大模型蓬勃发展的同时，也同样面临着诸多挑战。首先，虽然我国积极发展自主 AI 技术，但仍存在厂商产品兼容性差、缺乏统一标准的现象，需要推动国产 AI 生态不断完善。其次，大模型的迁移和部署过程相当复杂，需要专业的技术人员进行繁琐的配置和不断的调试。此外，AI 大模型在训练和推理过程中容易出现故障，需要快速有效的故障感知和定位机制来保障 AI 模型的稳定运行。

二、方案简介：

在这样的背景下，昆仑技术推出了 KunLun AI Space 大模型加速引擎，旨在为客户提供一站式的大模型迁移和开发解决方案。KunLun AI Space 不仅具备强大的硬件和软件支持能力，还提供了丰富的 AI 工具和服务，帮助客户快速完成环境部署、模型训练和模型应用，让 AI 大模型更好地服务于企业的业务发展。

图 1　KunLun AI Space 大模型加速引擎架构图

KunLun AI Space 大模型加速引擎是昆仑技术针对 AI 大模型落地过程中的各种挑战而推出的产品。昆仑技术将整个 AI 业务流程划分为三个阶段：点亮（环境部署）、跑起来（模型训练）和用得好（模型应用），并提供相应的工具和服务来支持每个阶段的工作。

点亮阶段（环境部署）

在点亮阶段，KunLun AI Space 通过昇腾使能加速工具/服务，为客户提供端到端的专业服务。昆仑技术利用深厚的操作系统和硬件调优全栈能力，结合自研的 EasyAscend 开局工具，帮助客户快速完成环境部署和调试。截至目前，昆仑技术已完成 5200+昇腾服务器的上架部署，得益于在环境部署方面的专业能力和丰富经验。

跑起来阶段（模型训练）

在模型训练阶段，KunLun AI Space 通过开发、重构算子，帮助客户解决硬件架构差异导致的训练问题。针对昇腾 AI 平台的特性，对算子进行深度优化，确保模型在迁移后能够正常跑起来。同时，昆仑技术还提供模型调优服务，解决精度和性能问题，提升模型训练效率。让客户在昇腾平台上可以轻松实现高效、稳定的模型训练。

用得好阶段（模型应用）

在模型应用阶段，KunLun AI Space 通过大模型故障感知定位套件实现故障自动感知、自动分析。昆仑技术积累了海量客户调优和故障解决的经验，并应用到产品特性上，确保 AI 模型在训练和推理过程中的稳定运行。此外，昆仑技术还提供了 AI 开发及应用服务，对 AI 计算资源实施统一分配调度，实现了 AI 模型从开发到推理部署的流程化。通过 KunLun AI Space 的专业服务，客户可轻松实现降低模型开发及应用门槛，缩短模型上市周期。

三、方案优势：

KunLun AI Space 大模型加速引擎在解决 AI 大模型落地过程中的挑战时，展现出了显著的优势：

1.全面性：KunLun AI Space 提供了一站式的解决方案，覆盖了从环境部署到模型训练再到模型应用的整个流程。为客户提供了丰富的 AI 工具和服务，满足了客户在不同阶段的需求。

2.高效性：KunLun AI Space 通过软硬件调优、算子开发和重构、模型优化等手段，显著提升了模型训练和应用的效率。让客户能够在更短的时间内完成 AI 模型的迁移和开发工作，更快地实现业务的稳定运行，提前为业务创造可观价值。

3.专业性：KunLun AI Space 拥有一支专业的 AI 调优团队，他们切实了解客户的业务诉求和痛点，为客户提供量身定制的解决方案服务，同时，在实践过程中不断积累丰富行业经验。此外，昆仑技术还提供了完善的售后服务和技术支持，确保客户在使用过程中良好体验。

4.经验丰富：KunLun AI Space 已经成功支持了 80+项目，解决了 400+大模型生态适配问题。客户涵盖了金融、互联网、运营商等重要行业，充分体现了 KunLun AI Space 在 AI 大模型领域的强大功能和特性。