智算平台运维运营技术研究报告

复旦大学
中国信息通信研究院云计算与大数据研究所
阿里云计算有限公司

2024年11月

编委会 / EDITOR

编委（排名不分先后）：

复旦大学：

吴力波、漆远、颜波、程远、韩丽妹、孙祥、张泰玮、李孟渚、张凯、葛治文、吴悠、关惠宇、黄岳、郭昕、蒋晨、徐跃东、林长龙、侯帅、江润丰

中国信息通信研究院云计算与大数据研究所：

栗蔚、马飞、苏越、赵伟博、桑柳

阿里云计算有限公司：

孙磊、付来文、李冬青、周昌盛、刘恩奇、王威、曹玉嘉、郎翊宇、杨仁远、张圣良

参编单位：

复旦大学

中国信息通信研究院云计算与大数据研究所

阿里云计算有限公司

版权声明 / Copyrig Notice

目录 / CONTENTS

前言 / FOREWORD

在数字化转型的浪潮中，智算中心扮演着越来越重要的角色，在国家数字经济和科技创新战略中的地位日益凸显。随着算力需求的不断攀升，智算中心不仅成为支撑人工智能、大数据、云计算等前沿技术发展的基石，更是推动经济社会发展的关键力量。

智算平台的运维运营是确保其高质量、稳定运行的关键。本研究报告基于复旦大学CFFF（Computing for the Future at Fudan）和阿里云智算中心的建设、运维、运营经验及中国信息通信研究院在此领域的研究成果，构建智算平台运维运营框架及评价体系。智算平台运维运营主要由三大能力域构成，一是AI运营，致力于人工智能模型的全生命周期管理，二是平台运营，着眼于提升用户体验和资源管理效率，三是平台运维，通过管理算力设备保障智算平台的业务连续性和系统安全。为客观衡量智算平台的运维运营水平，本报告从智算平台的基础设施、AI运营、平台运营和平台运维四个能力维度展开研究，提取通用、专用评估指标，构建智算平台运维运营评价体系，以期为行业内智算平台的建设、运维运营、能力评价提供参考。

智算平台运维运营是一个充满挑战的新兴领域，需要不断探索和创新。本研究报告旨在为业界提供更加全面、深入的研究视角，以促进智算平台运维运营的专业化、标准化和智能化发展。本研究报告仍有不足指出，期待业界专家和广大读者提出宝贵的意见和建议，共同推动智算平台运维运营领域的发展与完善。

01

研究背景及价值

算力的现状和发展趋势

智算平台的现状和发展趋势

智算平台的运维运营现状与面临挑战

研究背景及价值

1.1 算力的现状和发展趋势

随着数字化转型的深入和人工智能、大数据、云计算等新兴技术的广泛应用，算力已成为支撑经济社会发展的关键基础设施。中国作为全球第二大经济体和数字技术应用的前沿阵地，其算力需求呈现出爆发式增长态势。2024年政府工作报告中提出，大力推进现代化产业体系建设，加快发展新质生产力。要深入推进数字经济创新发展，制定支持数字经济高质量发展政策，积极推进数字产业化、产业数字化，促进数字技术和实体经济深度融合。深化大数据、人工智能等研发应用，开展“人工智能+”行动，打造具有国际竞争力的数字产业集群。实施制造业数字化转型行动，加快工业互联网规模化应用，推进服务业数字化，建设智慧城市、数字乡村。深入开展中小企业数字化赋能专项行动。支持平台企业在促进创新、增加就业、国际竞争中大显身手。健全数据基础制度，大力推动数据开发开放和流通使用。适度超前建设数字基础设施，加快形成全国一体化算力体系。我们要以广泛深刻的数字变革，赋能经济发展、丰富人民生活、提升社会治理现代化水平。

中国算力的快速发展为数字经济提供了强有力的支撑。随着“东数西算”工程的推进，中国的算力布局更加优化，特别是智能算力的快速增长，为中国在AI和大数据时代的增长提供基础。未来，中国将继续加强算力基础设施的建设，推动技术创新，完善政策和标准体系，构建全产业链生态，以促进算力产业的健康、高效和可持续发展。

1.2 智算平台的现状和发展趋势

本研究报告讨论的智算平台，是指通过使用大规模异构算力资源，用智能算力(GPU、FPGA、ASIC等)，主要为人工智能应用(如人工智能深度学习模型开发、模型训练和模型推理等场景)提供所需算力、数据和算法的设施。智算平台作为算力产业的重要设施，支撑着人工智能及相关产业的快速发展，但当前的智算平台多采用硬件驱动模式，存在水平较低、分割化严重、生态建设不足等问题，难以满足AI对“大数据、大计算、大模型”的需求。当前，美国等国家在算力、算法和数据方面已形成先发优势，而中国的公共智算平台及生态与之存在差距，特别是在AI公共算力设施及部分AI芯片上。AI算力及其服务市场可能出现“碎片化”，低水平、小规模的智算中心无法支撑大模型训练任务，可能导致资源浪费。面对上述等形势，国家和地方政府积极出台相关政策，推动智算平台的建设和算力产业的发展。

为了支持通用AI的发展，满足不同场景下的算力需求。智算平台将弥补传统计算中心的局限性，提供更广泛的服务，满足更多行业和领域的算力需求。此外，智算平台也通过优化算力资源配置、支持实时和离线计算需求等方式节约能源和成本。未来一段时间里，高性能算力产业生态体系也会是建设重点，推动产业链上下游协同发展，形成统一开放的AI算力产业生态。智算平台的发展可以降低中小企业的算力使用门槛，提升算力设施的普惠服务能力，加速赋能各行各业，推动产业数字化转型。智算平台正处于快速发展阶段，未来智算平台的建设也会是算力建设的重点，为算力的蓬勃发展提供平台服务基础。

智算平台离不开算力相关服务的专业化。算力相关服务作为智算平台的核心支柱，其发展必须与智算平台的整体进步相匹配，以确保整个系统的协调性和效能。算力服务的专业化不仅体现在技术层面的深度融合和智能化管理，更在于服务模式的创新、生态构建的完善以及安全合规的强化：通过结合LLMOps等思想，实现算力资源的智能调度和优化配置，提升服务效率和响应速度；探索按需服务、弹性服务等新型服务模式，以满足用户在多样化和个性化算力需求方面的期望，增强服务的灵活性和适应性；构建开放、共享的算力服务生态系统，促进跨行业、跨领域的协同创新和资源共享，以实现算力服务的可持续发展；加强算力服务的安全性和合规性，确保数据安全和用户隐私得到有效保护，构建用户信任的基石。

智算平台运维运营的价值

智算中心投资规模巨大，其能力与运营效率将成为运作的关键，构建合适的运维运营体系可有效地保持智算平台长期稳定运行，高效地管好和用好算力，并提供管理的实践，技术和工具的集合。

智算平台的运维围绕着模型服务，算力服务，容器服务，网络服务，存储服务以及安全服务等方面进行。智算平台的运营包含用户的日常管理及 AI 运营两个重点，用户运营包括用户管理、用户答疑、账单收费、工单管理、知识库建设等方面，AI 运营包括数据集运营、模型管理和部署、模型微调、提示词工程等能力。

智算运维运营平台为工程师提供了一个协作环境，该环境促进了数据和模型迭代探索、实时协作实验跟踪、提示词工程以及模型 Pipeline 的管理。同时，它还支持对大型语言模型（LLM）的控制模型转换、部署和监控。整体方案提供了一套完整的AI生命周期管理服务，从开发到部署再到维护，确保了平台的高效运行和持续优化。建设智算运维运营平台和相关团队，可以为平台带来如下保障：

1.确保服务连续性：
通过有效地运维运营，智算平台能够保证服务的连续性和稳定性，避免因故障或性能问题导致的服务中断，通过日常巡检和监控可以降低重大故障的发生概率。

2.提升用户体验：
良好的运维运营能够快速响应用户需求，提供及时的技术支持和问题解决方案，从而提升用户满意度。

3.研发效率提升：
通过工具研发的支持，智算运维运营平台允许团队更快地开发模型，提供更高质量的模型，并更快地部署到生产环境中。

4.优化资源利用：
通过精细化的资源管理和调度，可以提高计算资源的利用率，避免资源浪费，降低运营成本。

5.知识管理：
建设和维护知识库，促进使用方法和经验的共享，降低初学者的门槛。

6.模型微调、推理和监控：
模型微调优化模型以执行特定于领域的任务。模型推理可以基于现有知识管理生成内容。

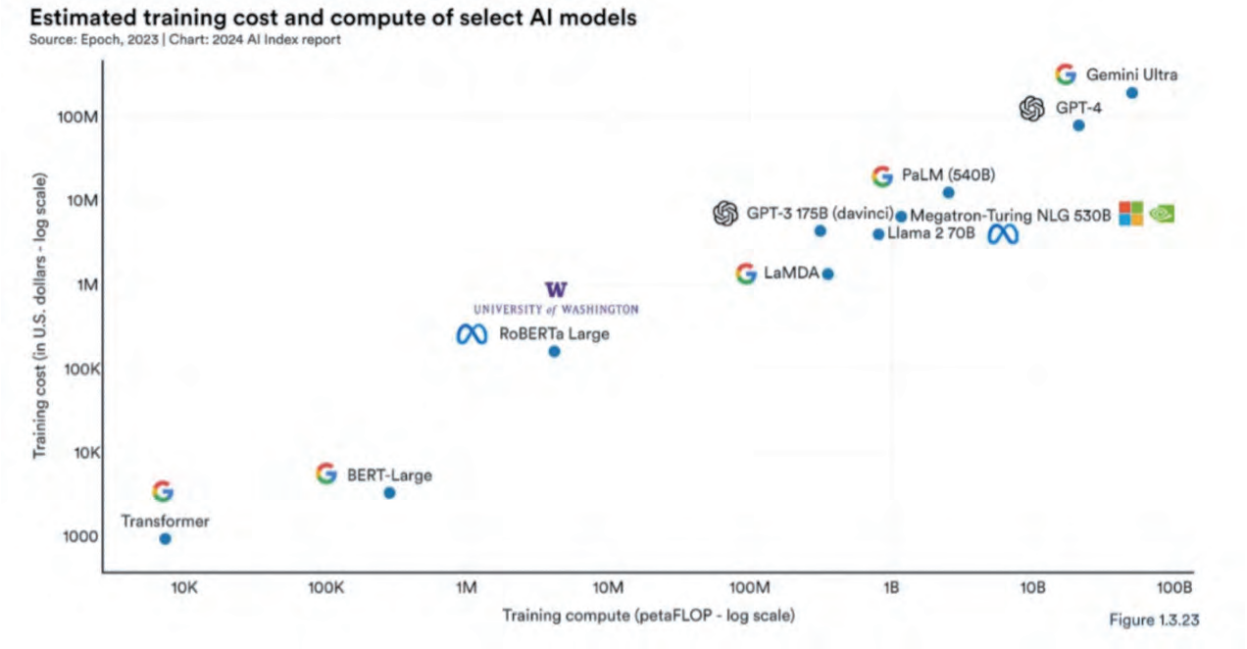
7.确保模型性能：
通过持续的监控和维护，智算运维运营可以确保模型在生产环境中的性能稳定，及时调整以适应新的数据和需求。

8.可扩展性：
随着业务需求的增长，智算运维运营支持平台的无缝扩展，可以灵活地增加计算和存储资源。

复杂的 AI 技术和智算生态对用户是一个挑战。高质量的智算平台的运营运维能力不光可以提升平台的稳定性，做好资源和用户管理，同时也降低 AI 模型的研发门槛，将研发好的 AI 模型快速应用到实际场景中。尤其对于那些工程能力相对薄弱的组织，如部分中小企业、具有 IT 诉求的非 IT 企业，智算平台的运维运营能力尤为关键。这些组织可能缺乏独立维护复杂 AI 平台的经验，依赖外部提供的高质量运维运营服务，可以加速创新孵化过程。

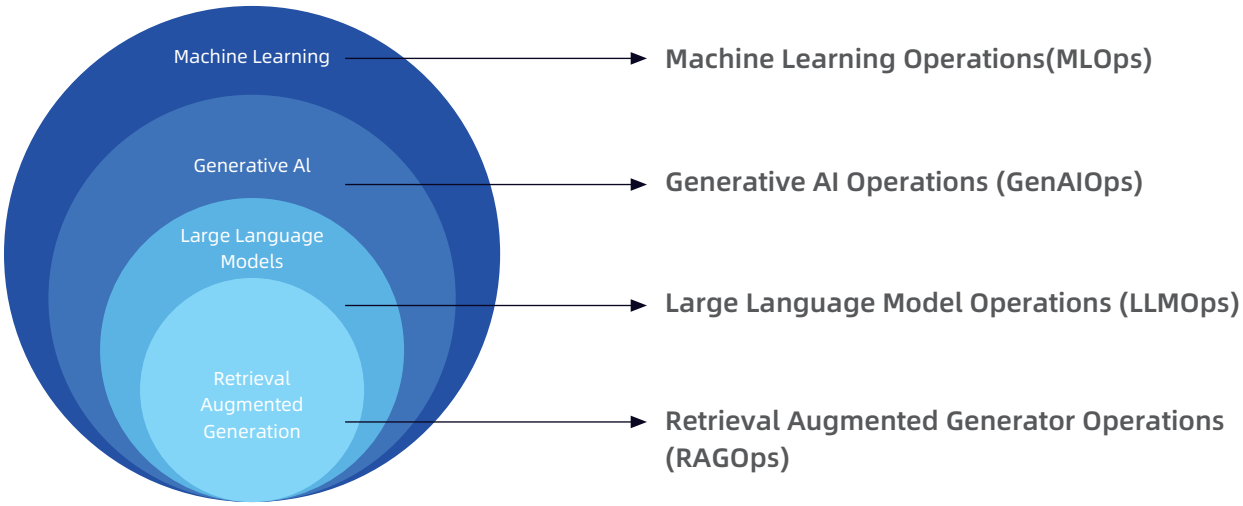
1.3 智算平台的运维运营现状与面临挑战

随着 AI 技术的发展，算力训练需求增长，智算设备紧缺，训练大型 AI 模型的成本变得极其高昂。OpenAI 的 CEO Sam Altman 曾透露，GPT-4 模型的训练成本超过了1亿美元。



▲ 图 1 大模型训练算力需求变化

国外在智算平台的建设和运维方面积累了丰富的技术和实践经验，有专业的团队负责智算平台的建设和运维工作。这些团队通常具备跨学科的知识和技能。目前，已经出现了 LLMOps 的概念，除了计算资源的管理和调度之外，还包括对 AI 模型全生命周期的管理能力。



▲ 图 2 AI领域不同层级运维理念

当前，国内智算平台运维运营相关领域的资料有限，尚未形成体系化的智算平台运维和运营解决方案。智算平台运维运营方面的不足，以及完善运维运营体系的必要性，主要体现在以下几个方面：

1.缺乏成熟的运维运营模式和经验。

2.缺少交叉学科的专业人才和团队。

3.需要建立更加完善的算力资源管理和调度机制。

4.运维运营缺乏对AI模型全生命周期管理的深入理解和实践。

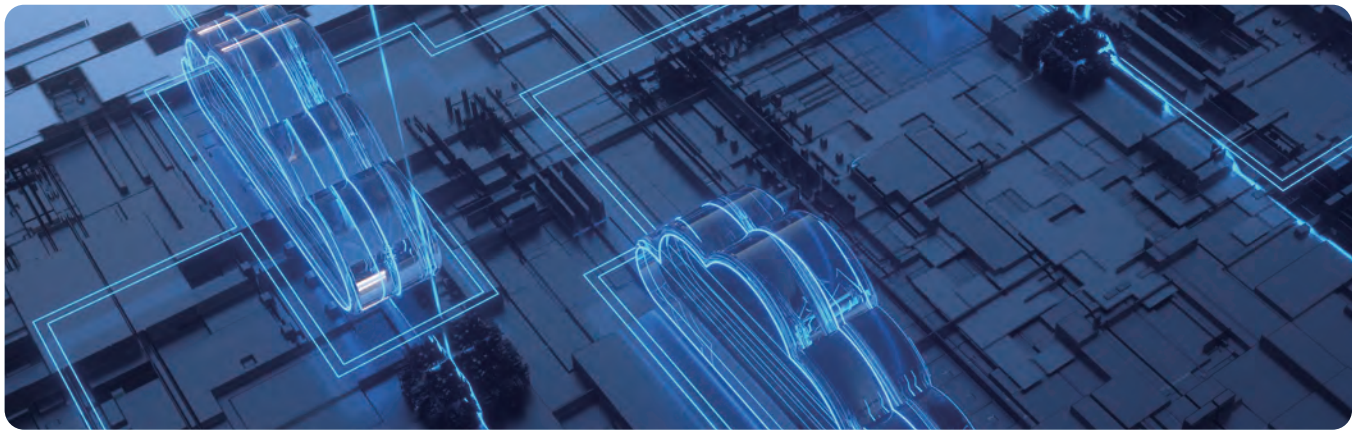
5.需要加强智算平台的安全性和稳定性。

随着用户的增加、算力供给增长以及服务生态的多样化，智算平台的运维和运营存在着较大的挑战，主要体现在人才缺失、流程和工具化能力缺乏、相关技术门槛高运营运维难度大、任务失败后排障困难等几个方面：

1.3.1 人才供给挑战

人才供给挑战主要体现在两方面，一是人才紧缺，缺乏具备必要专业知识和技能的人才，导致招聘难度增加；二是传统运维难度大，传统运维方式面临挑战和巨大的学习成本，缺乏高效的运维经验和标准。

传统运维运营方法与智算平台的运维运营要求之间存在较大差距，主要体现在AI模块的运维支持上。人工智能技术作为近几年的新兴的领域，综合了机器学习、深度学习、自然语言处理和计算机视觉等技术，对问题排查的人员能力要求很高，目前都由全栈型和有经验的算法工程师解决。



知识领域	传统运维运营人员	智算平台运维运营人员	备注
算力规模与性能			
CPU集群算力	1	1	集群传统高性能计算任务，需要运维机器
GPU集群算力	0	1	AI任务主要为各种GPU卡，需要运维机器和升级驱动
高性能存储	0	1	高性能存储，吞吐快，性能提升
平台服务能力			
预装软件数量	0	1	科研软件安装，外部工具接入
交互式环境	0	1	算法开发环境的使用和DEBUG
性能优化	0	1	AI任务和资源优化
任务调度	0	1	诊断K8s和Slurm调度的问题
运维与运营			
运维监控	1	1	帮助用户建设机器的运维体系，并且进行平台的变更操作
规章制度	1	1	帮助用户建机器使用制度
运营团队	0	1	协同机器的运营和平台的管理
AI和高性能计算			
HPC任务	0	1	HPC任务的使用
AI建模	0	1	模型的训练和推理
大语言模型框架	0	1	并行计算，PyTorch 和 TensorFlow 等框架
大数据	0	1	大数据加工，数据传输
算法开发与服务	0	1	协同用户解决算法训练的运维问题
AI和HPC镜像安装管理	0	1	为用户下载和安装镜像
其他			
网络性能	1	1	网络问题的诊断
容器化技术	1	1	POD的诊断，重启和删除
网络安全	1	1	网络安全能力建设

▲ 表1传统运维运营人员与智算平台运维运营人员能力对比　　注：1代表运维运营人员必备能力，0代表运维运营人员非必备能力

根据斯坦福大学 2024 年发布的《Artificial Intelligence Index Report》，智算领域的技术进步正在加速，但同时也带来了对专业人才和先进设备的巨大需求。而能够理解 AI 又愿意做智算运维运营平台的人员非常稀缺，招聘难度巨大。

Top 10 specialized skills in 2023 AI job postings in the United States， 2011-13 vs. 2023

Source： Lightcast，2023Chart： 2024 AI Index report

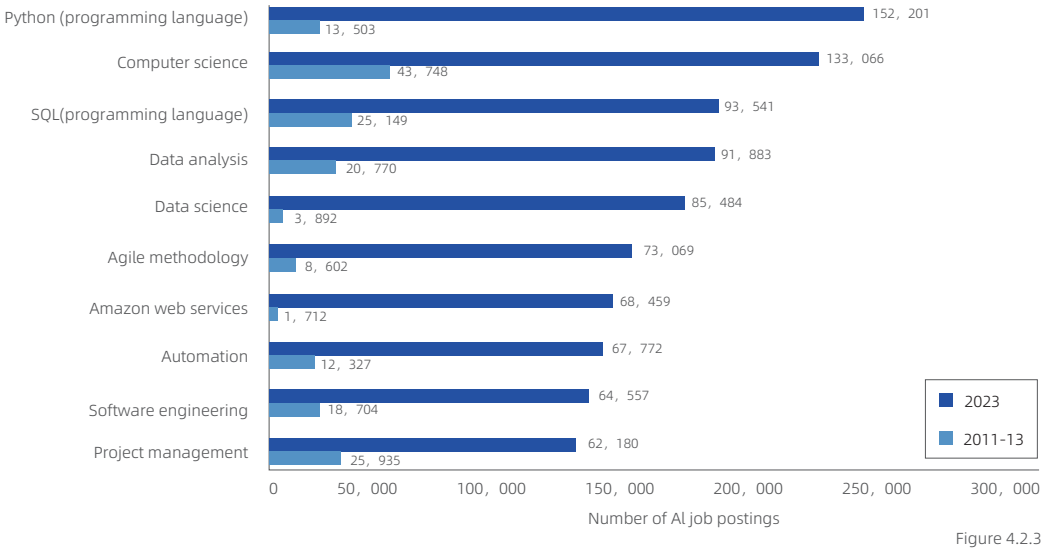


Figure 4.2.3

▲ 图 3 2023年美国AI从业人员top10工作技能具备人数与十年前对比

随着技术人才使用的智算平台设备日益昂贵，对运维运营人员的要求也相应提高。不仅要求他们掌握高水平的专业技术能力，更要具备出色的管理与决策技能，以保障智算平台的高效运行和持续创新。当前国内在这一领域面临运维运营人才短缺的问题，亟需在持续的教育和实践中培养。合适的智算平台运维运营人员，不仅要有传统运维运营的基础，还要对人工智能技术有深刻理解，掌握相关的管理和决策知识，以适应智算平台在数字化转型和AI升级中的新需求。

1.3.2 流程和工具化能力缺乏

目前，大模型训练的生态系统仍在建设之中，相关的流程和工具尚未完全产品化。同时，我们还缺乏统一的标准和接口来管理相关资源。例如，对于模型的运行状态、对应的 GPU 机器以及平台稳定性，我们还需要一个统一的监控和统计系统；大规模 GPU 集群的扫描软件、AI 训练生态系统，推理和模型输出等都处于创新阶段。偏定制化的需求，面临流程缺失和工具缺乏等问题，极大程度地增加了运维运营工作的难度，目前市面上类似 Datadog、HuggingFace、atabricks 等公司都在积极地解决 AI 任务监控和训练的生态问题，未来有望可以标准化输出。

1.3.3 智算门槛高，运营运维难度大

目前智算的高技术门槛和运营运维的复杂性使得许多企业和研究机构望而却步，其主要原因在于对 GPU 资源的大规模依赖。此外，智算系统的设计和实现需要跨学科的知识和技能，包括机器学习、数据科学、软件工程等，均成为了运维运营工作开展的挑战。

在运营和维护智算系统时，团队面临的挑战尤为严峻。系统稳定性的维护需要持续地监控和及时故障排除，而性能优化则要求对系统架构有深入的理解。随着技术的快速发展，智算系统需要不断地更新和升级，以适应更大规模的算法参数和更大的数据集,通过更敏捷的模型应用部署平台，来满足 AI 模型对实际业务场景的适配。为了克服这些挑战，企业和研究机构需要投入更多的资源进行人才培养、技术研发，并探索和总结更高效的运维和运营策略。

1.3.4 任务失败后排障困难

计算任务失败原因分析路线非常复杂，从硬件到上层框架链路长，涉及的领域众多，对目前运维运营人员的技术要求较高。任务排障困难体现为如下几方面：

1.系统架构复杂：
智算平台通常由多个模块组成，如底层基础设施、机器学习平台和运维运营平台等，每个模块都有其特定的功能和架构，问题定位困难。

2.硬件和软件问题：
底层硬件问题（如ECC错误、NVLink错误）和软件配置问题（如Shell启动失败、缺少配置文件）可能影响系统运行，需要专业知识进行诊断。任务调度失败、训练速度慢、资源不足（如OOM错误）等问题，需要对平台执行 AI 任务的逻辑有一定了解。

3.用户权限和资源管理：
用户权限设置、资源申请、工作空间配置等方面的问题，需要对平台的运营体系有深入了解才能解决。

4.环境配置和依赖问题：
AI 模型训练环境配置复杂，涉及镜像、数据集、代码等 AI 资产的管理，以及依赖包的安装和配置问题。

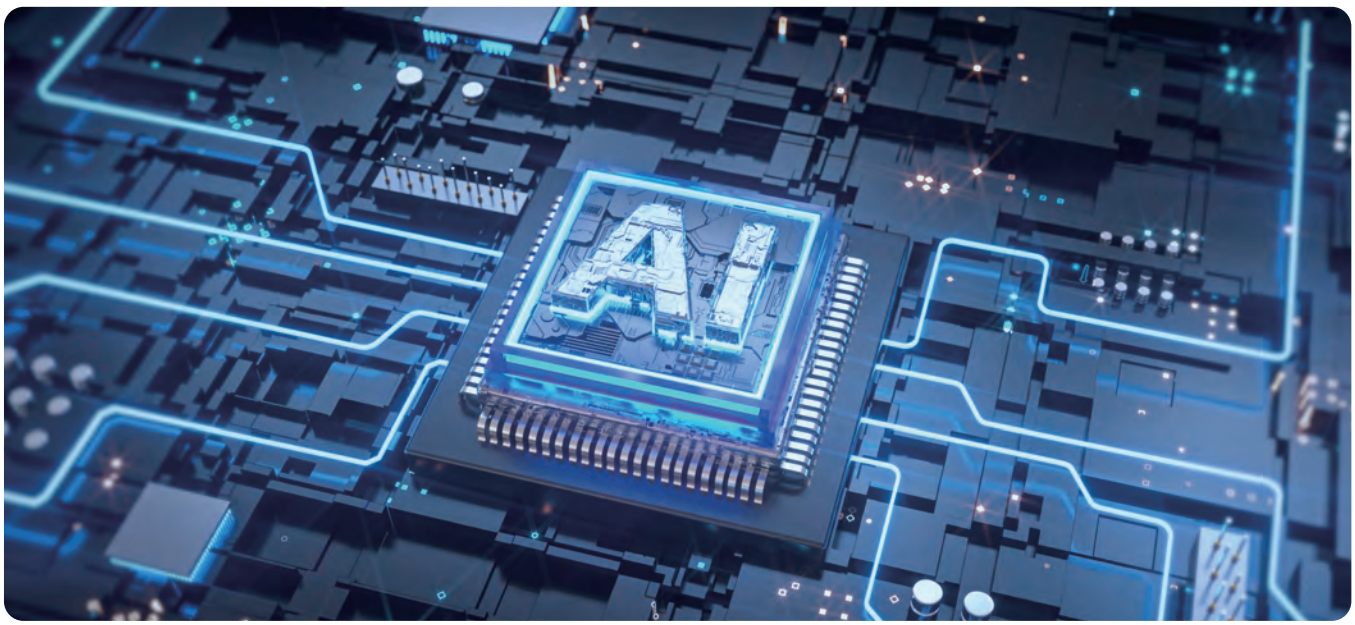
5.网络和存储问题：
网络连接问题、存储设置错误、文件操作限制等，均可能影响用户的正常使用。

6.硬件故障：
GPU 卡顿、掉卡、硬件损坏等问题，需要硬件维护团队及时介入。

7.用户熟悉度不足：
用户对平台的使用不熟悉，导致操作错误或无法充分利用平台功能。

智算平台的任务排查是一项极具挑战的工作，它要求运维人员不仅要有深厚的技术背景，还需对整个系统架构有全面的理解。从底层硬件的稳定性到软件配置的精确性，每一个环节都可能导致训练任务执行失败。同时新的挑战不断涌现，如确保数据安全、遵守合规性要求、处理大规模并发请求等，都进一步增加了任务排查的难度。

根据目前智算平台运维运营的现状，为了提高智算平台的运维效率和稳定性，需要完善自动化监控和故障排除工具，加强人才培养，确保智算平台在面对日益复杂的AI任务时，仍能保持高效和稳定，并且将大模型等AI技术有效得应用。本研究报告面向智算平台支持AI模型训练的全生命周期，总结当前智算平台的运维和运营难点，并提出了相应的解决方案。



02

智算平台运维运营

智算平台运维运营中心主要功能

智算平台运维运营组织架构及制度体系

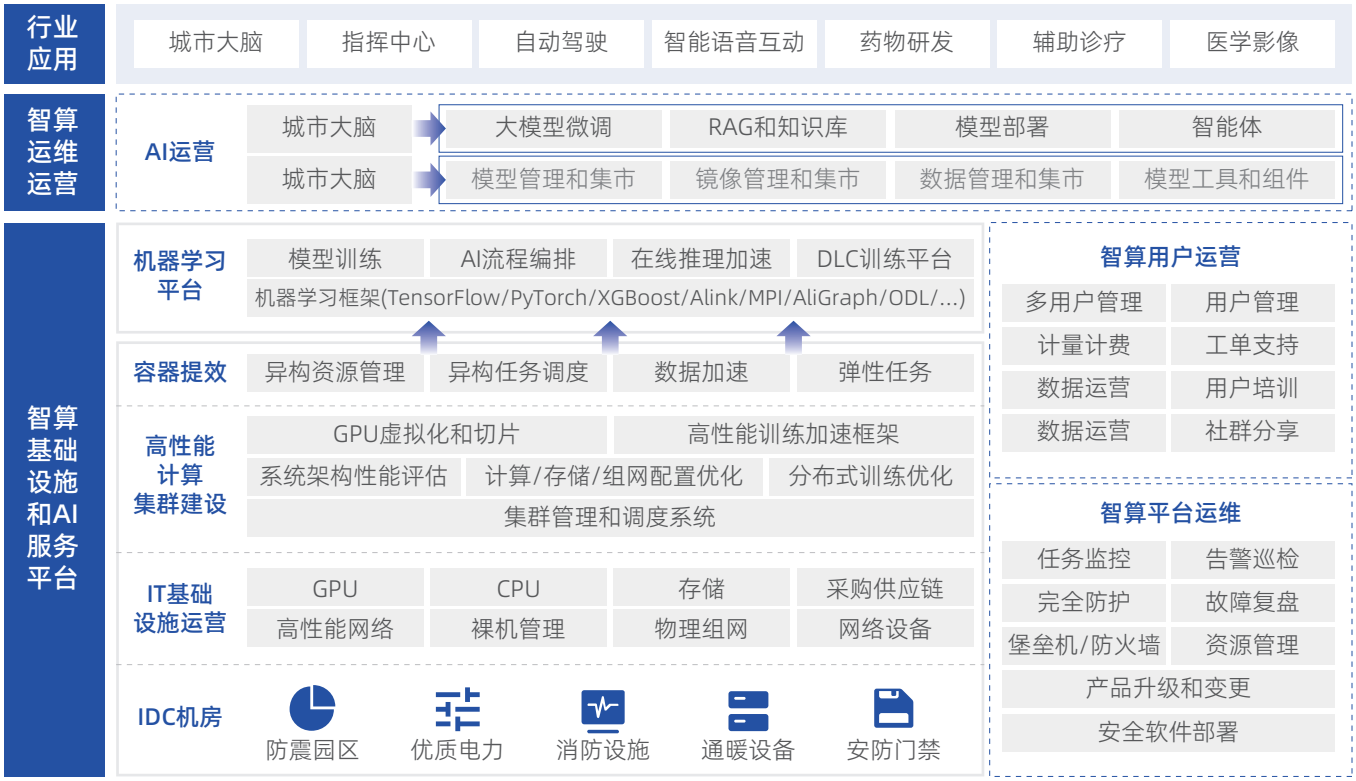
AI运营

智算平台运营

智算平台运维

2. 智算平台运维运营

智算平台为支持多样的行业应用而建设，智算平台运维运营在智算平台体系结构中的位置如下图所示：



▲ 图 4 智算平台运维运营体系结构

智算基础设施和AI服务平台位于智算平台体系结构的最底层，主要提供两个重点能力：基础设施 IaaS 和 AI 平台 PaaS。用户无需组建和运维复杂的 GPU 机器、存储和 RoCE 网络，即可使用高拓展性、高性能的 IaaS+PaaS 的环境：

1.基础设施IaaS：

IDC 机房、网络交换机（RDMA网络交换机、通用网络交换机）、算力服务器（智算算力服务器、通用计算服务器）、存储服务器等能力；同时还有基于基础设施的集群建设，为上层平台和应用提供算力、存储、网络、容器、容器镜像、安全等服务。

2.AI平台PaaS：

提供随开即用的的 AI 作业平台，完成对 AI 模型（包括大模型）的开发和训练。

智算运维运营在智算基础设施服务平台和行业 AI 场景之间起到桥梁作用，除了为用户提供统一的资源管理和算力资源的监控，也为上层的智算模型运营提供产品和服务（模型微调、Agent、AI资产生态运营等），有效地提升智算平台整体的性能和用户体验。

通常智算运维运营能力由智算平台运维运营中心承载，其能力进一步细分为 AI 运营、平台运营及平台运维。

2.1 智算平台运维运营中心

▲ 图 5 智算中心运维运营总览

智算运营运维中心主要分为三个重点的模块：

2.1.1 AI运营

AI 模型的开发，尤其是大语言模型的开发过程包含许多复杂组件，如数据加工、数据预处理、提示词工程、模型微调、模型部署、模型监控等，同时还需要跨团队的协作和交接，从数据工程到数据科学再到机器学习工程，整体流程需要严谨运营以及相对应的产品工具。AI 运营的目标是通过产品工具和专家服务，降低用户在 AI 模型训练和应用的工程门槛，提高大模型应用的开发效率。

AI 运营非常重要，包括可视化，透明度和可解释性。通过AI模型的运营模块，可以让非技术人员参与到 AI 应用的运作中。其中主要包含模型运营和 AI 资产运营。

1.模型运营：

模型运营的目标是为了释放大模型的价值，其中包含模型微调、提示词工程、智能体（包含各种工程组件）以及模型监控等能力，同时也包含大模型专家服务用来解决在模型训练和推理过程中遇到的问题。

2.AI资产运营：

主要面向丰富的 AI 资产生态，具体包含：

1) **模型集市：**包含官方开源的大模型和组织内公开的大模型，可以进行模型版本的控制更新，分享和部署。
2) **数据集市：**包含官方开源的数据集，和组织内公开的数据集，可以协同开展数据上云，数据加工，数据共享等。
3) **镜像集市：**主要包含支持各种大模型的不同镜像，来自不同的社区。
4) **实验集市：**主要包含各种业务组件，用于降低模型部署或者数据加工的工程化门槛。

2.1.2 平台运营

平台运营可以帮助企业利用已有的算力资产，向租户出售算力产品和增值服务，帮助用户更高效地使用算力。同时平台运营会有效地处理用户资源数据，给企业组织提供决策和实现平台，从而提高整体智算平台的运营效率，降低管理和维护成本。

1.用户运营：

用户运营主要包含用户权限管理、工单答疑、用户培训等。通过工单服务解决用户找人难、上手难、排查难的痛点。通过智算知识库帮助管理者在运营过程中持续沉淀智算行业的宝贵经验。

2.资源运营：

一站式的资源全生命周期管理。资源运营主要包含全面的资源管理，包含不同类型、不同收费模式和计算资源进行混合管理。用户能够在平台对计算资源进行从申请、审批、创建、变更到回收的全链路管理动作，并且平台能够精确记录资源的申请或变更记录、资源的项目归属和资源的计费主体。

3.运营管理：

包含管理经验的运营流程设计、数字化管理、经营分析和计量计费等模块，帮助用户高效、便捷地对智算场景开展更全面精细和准确的运营。通过数字化管理和经营分析可以快速的发现问题，提升用户体验和资源利用率。

2.1.3 平台运维

通过端到端地对物理资源、机器学习平台及上层应用进行日志采集和监控，平台运维能够快速且精确地诊断问题，迅速响应并预防重大问题的发生。同时，平台运维提供专门针对智能计算任务的运维服务，以解决用户在使用时硬件基础设施时遇到的功能和性能等问题。

1.业务连续性：

业务连续性需要软硬一体的全栈运维来支持，覆盖了从各个型号的 GPU、CPU 硬件、并行存储节点，以及网络和通信等底层基础设施硬件。此外，还需支持上层的容器服务，确保容器和容器间的通信，以及每个容器里代码平稳地运行，从而产生可靠的 AI 运算结果。在业务连续性方面，需要全面的日志采集、业务监控、故障应急、变更升级和 AI 运维巡检能力，为整个 AI 系统提供高效地运行支持。

2.安全防护：

安全体系的设计需要多个重要的参与方，运维团队需要跟安全团队紧密合作，确保技术基础设施的可靠性和安全性。运维团队负责日常系统维护、软件部署和故障排除，安全团队则专注于评估风险、监控威胁和强化防护措施。

3.智算运维：

智算运维模块不同于传统运维的服务能力，主要针对大模型训练和推理的相关业务需求开展性能分析优化、算力和存储扩容、软件镜像安装、模型训练报错诊断、大模型迁移GPU卡等比较新兴的运维服务能力。

2.2 智算平台运维运营组织架构及制度体系

2.2.1 组织架构

为保障智算平台的安全稳定和业务的长效运营，平台运营运维需要如下组织构成。

1.智算平台运营组：

提供资源受理和办理、资源账单服务、工单受理、赋能培训、产品需求缺陷管理、解决方案服务、资源目录梳理、资源开通流程规范、资源计量计费规范、资源效能规范等服务。

2.智算运维保障组：

保障平台软硬件稳定性服务、服务完成平台日常变更、告警处理等问题处理，人员通常配置驻场运维服务 5*8，远程运维保障服务 7*24。

3.AI 应用运营组：

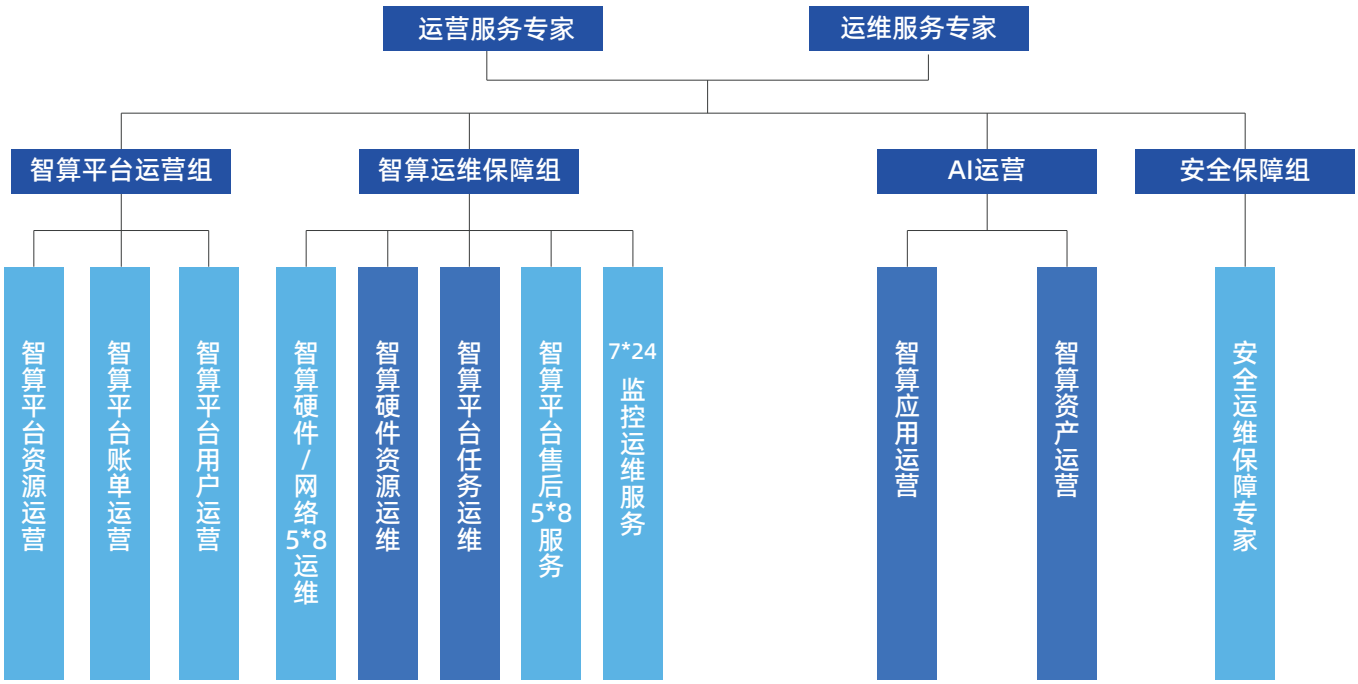
大模型模型运营负责提供模型部署的数据支持，确保模型可以稳定高效地推理和应用，同时确保用户关于 AI 模型的微调、RAG、Agent 建设可以顺利开展。

4.AI 资产运营组：

针对大模型开发过程中主要的生产资料数据集，模型和镜像进行有效的运营管理。数据运营负责收集、整理和存储用于大模型训练的数据集，确保数据的质量和完整性。同时负责为相关的模型提供镜像的下载和部署服务。

5.安全保障组：

负责平台安全架构设计，包括防御、监测体系构建，潜在风险识别和安全策略制定。



▲ 图 6 智算平台运营运维组织架构

2.2.2 制度体系

为了保障平台建设和运维运营过程中的整体稳定性和线上业务的正常运行，结合人员和工具的能力建立流程和问题管理机制。

1.资源管理：
建立资源分配和调度的规则，确保 AI 模型训练和推理任务能够高效利用计算资源。

2.故障恢复：
制定故障恢复流程，包括自动故障转移、备份和恢复机制，以最小化系统停机时间。

3.性能监控：
实施实时监控系统，跟踪集群的性能指标，如负载、响应时间、错误率等，以便及时发现并解决问题。

4.资源巡检机制：
定期进行资源巡检，确保资源配置得当，及时发现资源使用中的瓶颈和浪费问题。

5.用户管理：
建立用户管理体系，确保用户权限的合理分配，优化用户体验，包含用户在项目申请、账单结算、工单提出等多个环节的管理。

6.数据管理：
制定数据管理政策，确保数据的完整性、可用性和合规性，提高数据的质量和分析能力。

7.AI模型管理：
建立 AI 模型的全生命周期管理流程，包括模型的开发、测试、部署、监控和下线。

8.AI应用管理：
对 AI 应用进行系统化管理，确保应用的性能、安全性和用户满意度。

9.文档和知识管理：
维护详细的文档体系，记录系统架构、操作流程和故障处理案例。

10.成本管理：
监控和优化集群的运营成本，包括硬件投资、能源消耗和维护费用。

安全架构设计：

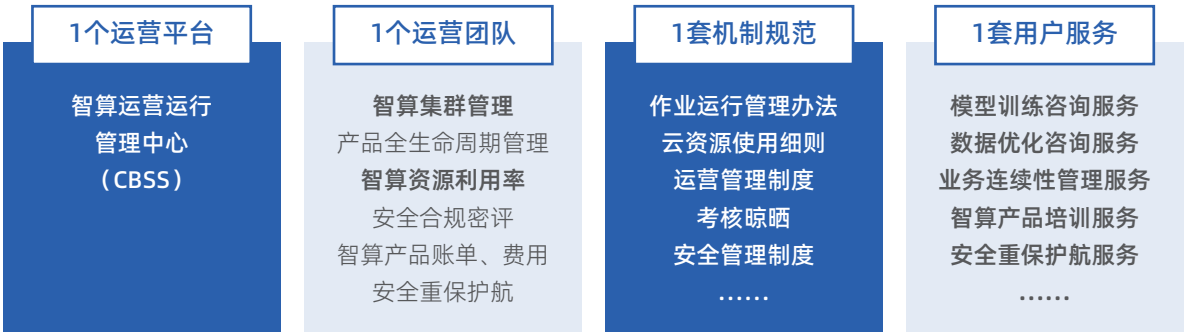
制定严格的安全政策和协议，包括访问控制、数据加密和网络安全措施，保护集群免受内外部威胁。

安全合规性和审计：

确保所有操作符合法律法规要求，并定期进行内部和外部审计。

产研协同体系：

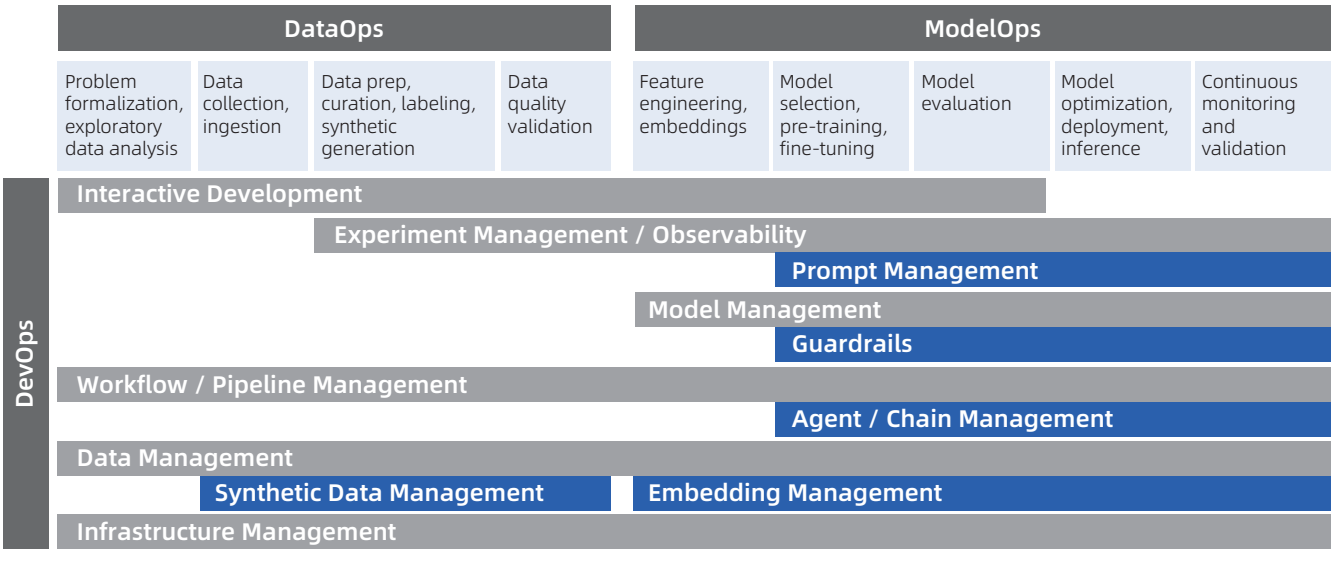
确保产品的缺陷和需求可以伴随着业务的发展快速迭代。对日常运维运营中发现的缺陷和需求可以快速解决。



▲ 图 7 智算平台运营运维制度体系

2.3 AI运营

AI 的运营主要包含模型运营和 AI 资产运营，其中模型运营主要为了完成 AI 模型的业务应用，AI 资产运营主要是为模型训练和推理提供高质量的素材。

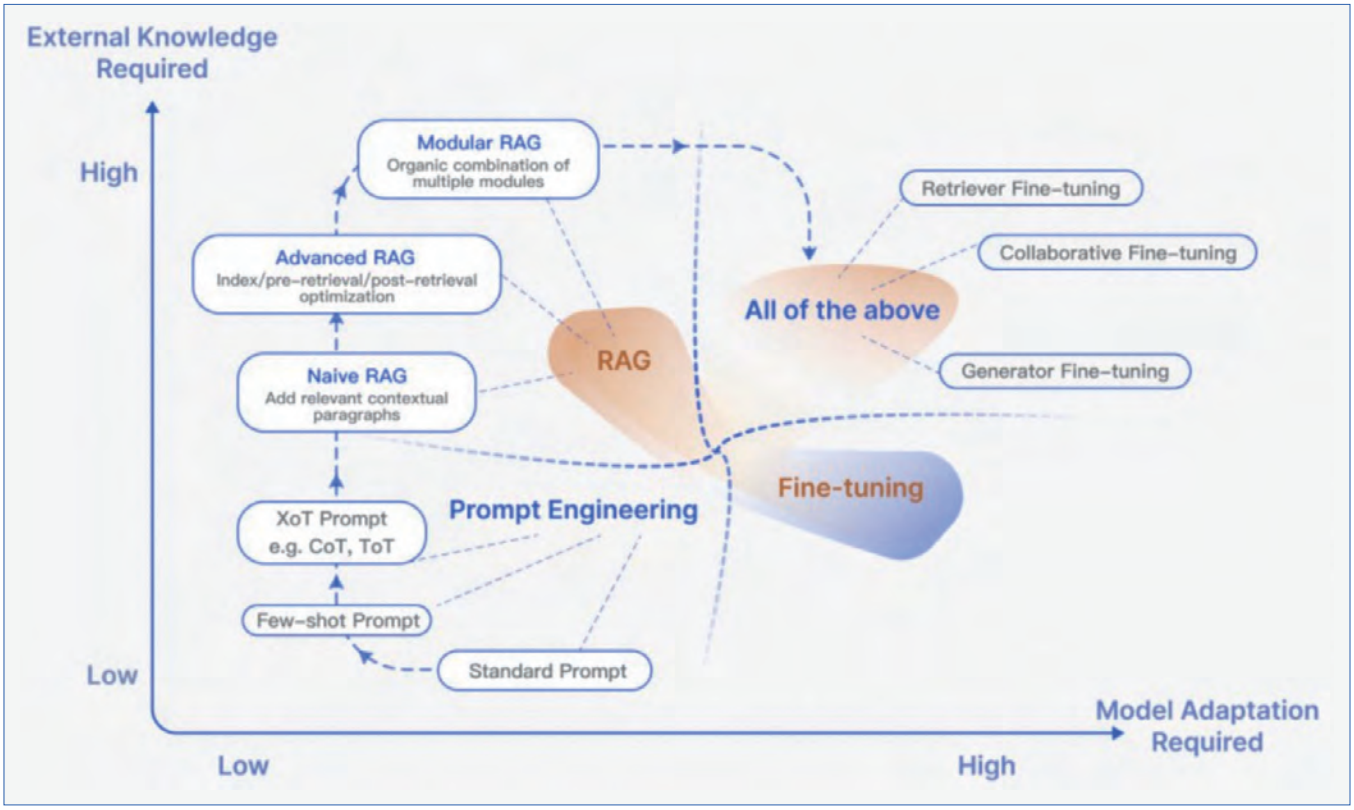


▲ 图 8 AI 运营范围划分

大模型应用还面临诸多挑战，例如开发团队还未适应大模型编程的需求，对大模型的实际应用场景理解、工具的选择（例如中间件、向量数据库等）以及团队的协作模式、如何构建 Prompt 等方面都存在一定的认知偏差。开发团队需要在大模型技术栈方面建立更多的共识，对于如何使用 RAG（Retrieval Augmented Generation）或者微调等应该有更明确的工作流程。

2.3.1 模型运营

模型运营是指通过大模型提供应用和服务。模型运营基于对外部的知识库的输入和模型是否需要调参可以分为微调、RAG 和提示词工程，帮助大模型快速回答专业领域的问题。另外智能体平台提供产品化工具，简化工程化能力，帮助用户快速部署模型和实现模型在实际业务场景中的价值。



▲ 图 9 微调、提示词工程、RAG 技术

2.3.2 模型微调 (Fine-tuning)

模型微调是指在预训练模型的基础上，针对特定的应用场景或数据集进行进一步训练的过程。可以通过预训练的 LLM 作为起点，然后在特定任务或领域的标记数据集上训练完成。这样做可以使得模型更好地适应特定的任务，提高其在该任务上的表现，其主要技术包括：

1.全微调：

用预训练模型作为初始化权重，在特定数据集上继续训练，全部参数都更新的方法。

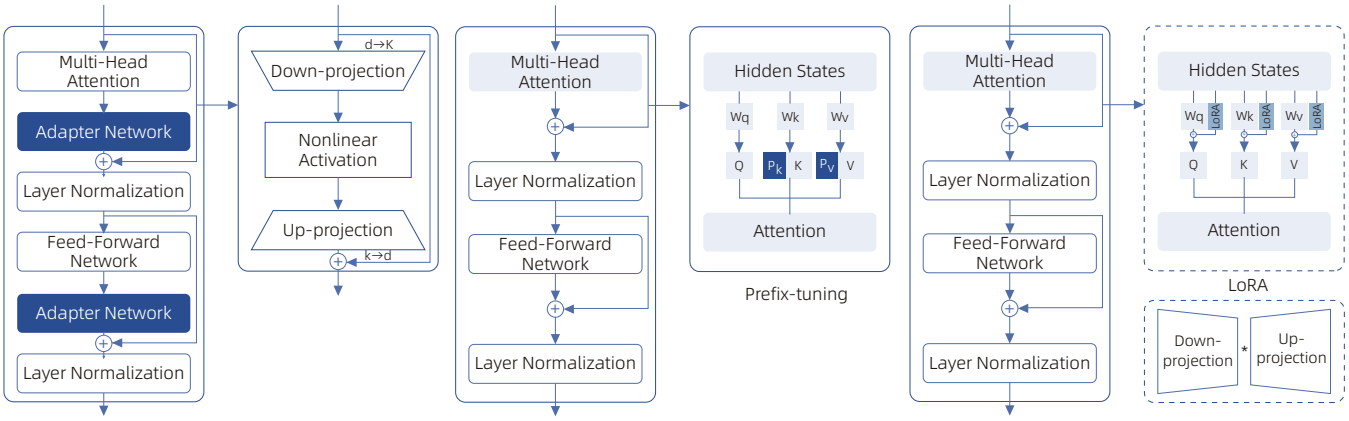
2.高效参数微调：

a) 增加额外的参数（Addition-Based）：Prefix Tuning、Prompt Tuning、Adapter Tuning。

b) 选取一部分参数的更新（Selection-Based）：BitFit。

c) 引入重参数化 Reparameterization-Based：LoRA。

d) 混合高效微调：MAM Adapter、UniPELT。



▲ 图 10 大模型参数微调原理示意图

2.3.3 RAG (Retrieval-Augmented Generation)

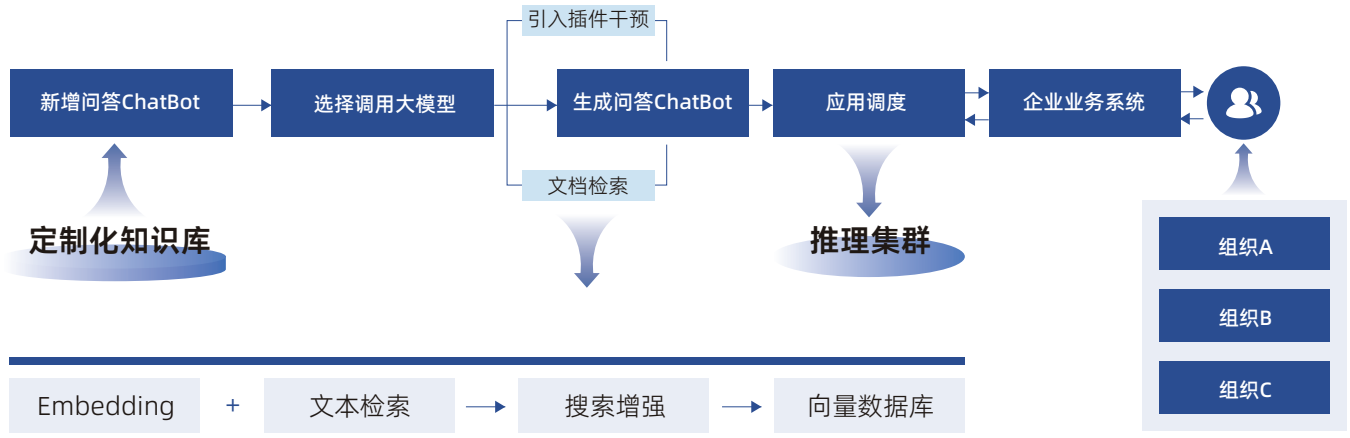
RAG 是一种结合了检索（Retrieval）和生成（Generation）的模型架构，它首先从一个大型的数据库中检索相关信息，然后将这些信息整合到生成模型中，以生成更加丰富和准确的输出。该方法有非常多的优势，例如：

a) RAG 通过将答案与外部知识联系起来，减少语言模型中的幻觉问题，并使生成的回答更加准确可靠。

b) 使用检索技术可以识别最新信息。保持了响应的及时性和准确性。

c) 透明度，通过引用来源，验证答案的准确性，增加对模型输出的可解释性。

d) 安全和隐私管理，RAG 凭借其在数据库中内置的角色和安全控制，可以更好地控制数据使用。



▲ 图 11 RAG的技术架构示意图

整体 RAG 系统包含两个阶段：检索阶段（Retrieval Phase）和生成阶段。其中在检索阶段，根据用户提出的问题，检索系统搜索用户上传的知识库，（该知识库可能包含文档、网页或其他形式的数据。同时知识库会被切成不同的片段以向量的方式存在向量库）。语言模型会把检索到的文档作为输入，结合问题和用户的原始问题，生成答案输出。

2.3.4 Prompt 提示词

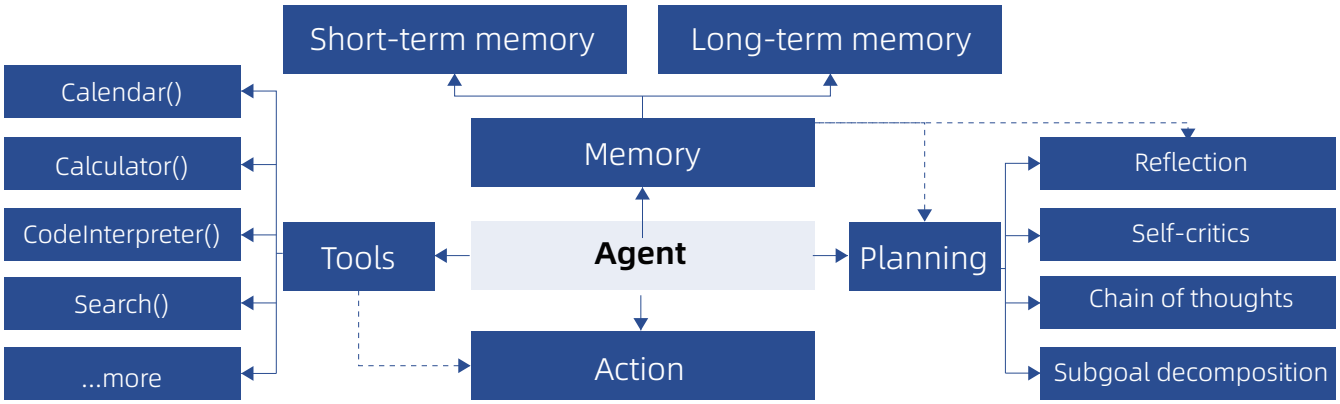
提示词是一种引导模型生成特定类型回答的方法。在一些生成模型中，通过精心设计的提示词可以引导模型生成更加相关和高质量的内容。高质量的提示词可以提升答案生成的质量，实现特定任务和目标，同时设定个性化的风格来适应多样化的需求。

2.3.5 智能体平台

智能体是人工智能领域的重要概念，它可以被定义为一个实体，可以在所处的环境中感知信息，并且根据这些信息作出决策，然后实现特定的目标和任务。智能体有自主性，感知能力和决策能力。其中 LLM 作为智能体的大脑，辅以规划、记忆和工具使用等关键组件。智能体能够将大型任务分解为更小的子目标，并使用短期和长期记忆来处理信息。

智能体平台通常指的是一个允许用户创建自己 Agent 的平台，这样的平台可以提供一系列的工具和服务，帮助用户快速构建和实现AI应用。

AI Agent 被公认为 AIGC 发展的确定性方向。大语言模型目前更加擅长对话和绘画类任务，导致 LLM 能力普遍存在固化。同时由于算力问题，LLM 的记忆里有一定的限制。而 AI Agent 有规划能力，可以将大任务拆解为自任务，并且可以自动化使用和调用工具，为大语言模型的应用带来了更广阔的空间。



▲ 图 12 AI Agent架构示意图

2.3.6 AI资产运营

AI 资产通常指的是在人工智能领域中，由企业或个人所拥有的、能够产生价值的资源或技术。第一步是 AI 资产的管理，确保算力可以有效利用，快速训练出符合业务场景需求的模型的关键。第二步通过 AI 资产的运营，让 AI 资产实现共享，可以轻松下载和训练组织之间的预训练模型，大幅节约研发者的模型训练成本和时间，构造组织内部 AI 的开源社区。这些资产包括但不限于以下几种类型：

1.数据集市：

数据是训练模型的基础。高质量的数据集可以提高 AI 系统的性能和准确性，且大模型数据庞大，开源数据集需要被登记和管理，平台需要实现组织内部的数据共享，为大模型训练提供语料库。该模块具备如下的能力：

a) 数据上云：

· 模块支持网站数据、专业文献、行业数据等多种安全数据源连续地导入数据。

· 支持对各种原始格式的数据格式，例如 PDF，DOCX，XML 接入平台，对结构化和非结构化的数据导入，实现大规模和大体量的数据上云。

b) 数据管理：

· 对数据的来源和权限，以及元数据属性进行管理。

· 对数据的业务属性和 AI 属性进行管理，例如数据集的领域，应用场景和相关权限。

· 对数据集提供上架，更新和下载的能力，对数据集全生命周期管理。

c) 数据加工：

· 平台提供以 Python 和 SQL 为基础的数据加工能力，提供特征工程服务。

d) 数据标注：

· 平台提供集成的标注工具，支持不同类型数据的标注需求，如图像、文本、音频和视频。提供直观的用户界面，使标注人员可以轻松地对数据进行分类和标记。

· 设计和实施标准化的标注流程，确保数据标注的一致性和准确性。且支持多人协作标注，实现标注任务的分配、审核和质量控制。

· 建立标注结果的反馈机制，允许标注人员根据模型训练的反馈调整标注策略。

· 自动化标注，利用机器学习技术，开发自动化标注工具，以减少人工标注的工作量。

e) 数据展示：

· 平台提供基础的BI报表建设能力，允许对数据集的相关结构化信息开展业务分析，对核心的数据指标进行可视化报表展示。

2.模型集市：

模型集市支持用户发布和下载开源的预训练模型。实现对模型共享和快速模型的部署。同时将用户训练好的模型进行上架、更新、版本管理，实现对模型的全生命周期管理。

a) 模型注册：

· 模型注册提供模型的上架能力，对模型的版本进行控制，快速完成模型的业务打标，如来源、应用场景、描述说明等。可以让用户快速找到对应的模型并且完成应用授权。

b) 模型部署：

· 模型部署是快速地将上架的模型部署到GPU计算资源，涉及到模型的容器化和推理权限配置等。对模型在实际应用中的性能和可靠性进行测试。

c) 模型库管理：

· 模型库提供一系列的开源预训练模型，这些模型可以针对图像识别、自然语言处理、推荐系统等进行优化。包含一系列的模型文档，用户反馈系统，调用次数和下载次数的监控。

3.镜像集市：

提供丰富的镜像资源库，允许用户浏览，选择和下载各种大模型训练需要的环境依赖。该模块通过提供预配置的镜像，显著简化了大模型训练和部署的复杂性，降低了工程实施的门槛，支持开发者在高效、可靠和安全的环境下进行大模型的开发和创新。

a) 镜像导入：

· 需要为官方镜像导入和用户镜像导入提供标准和验证流程，确保镜的质量和安全性。同时制定一套镜像命名规范、标准化镜像上架流程。

b) 镜像库管理：

· 建立流程服务和管理镜像库，包含关键镜像的官方源更新和软件更新，同时允许用户镜像访问权限控制和共享用户镜像。

c) 镜像诊断：

· 对于大型和复杂的模型，提供镜像诊断工具，帮助用户排查和解决镜像使用中的问题，并提供核心技术支持。

4.实验集市：

实验集市为研究人员提供了一个平台，用于管理、共享和协作实验流程和结果。基于算子和工具实现不同场景的算法业务流，对实验设计、执行、结果分析和共享。

a) 工具管理：

· 提供一个集成的工具管理平台，允许研究人员访问和管理各种实验工具和软件。同时支持工具的版本控制和依赖管理，确保实验的可重复性。

b) 实验管理：

· 实现实验的全生命周期管理，从实验设计、执行到结果分析，支持实验的自动化执行。

c) 实验工具研发：

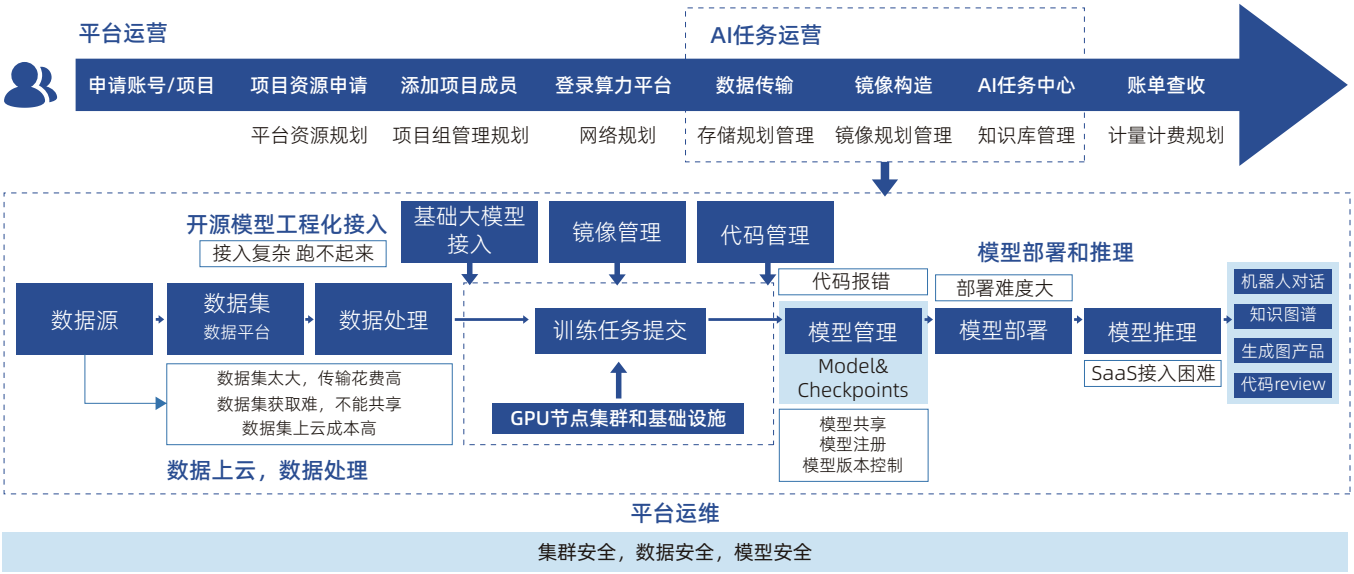
· 鼓励和支持研究人员开发新的实验工具，以满足特定的研究需求。提供工具开发的资源和指导，促进创新和协作。

2.4 智算平台运营

智算平台的运营从用户的使用需求开始，覆盖资源使用的全流程，形成智算平台运维运营体系，由用户运营、资源运营和运营管理三个方面构成。

2.4.1 用户运营

对智算平台运营的范围和内容从沿着用户使用路径展开，具体包括AI常见问题技术支持、知识库、培训以及费用管理和账单查收等方面。



▲ 图13 智算平台用户运营流程示意图

1.用户管理

通过产品化的能力以及相关的运营运维流程，为用户提供一系列针对智算平台使用的服务，让用户可以高效地管理自己的账户、资源和服务。同时运营团队需要制定一系列的规范和机制，指导用户高效地使用算力。其中包含：

1.用户和项目组注册：为用户提供账户的注册、项目组注册和管理等能力。

2.资源开通：为用户提供资源和规格的选择、开通算力和存储资源。

3.订单管理：用户可以管理账户的资源订单，以及上传和编辑合同模版。

4.工单管理：用户可以提交和跟踪工单，以及查看故障待办和当前进展。

5.账户资金管理：用户可以充值账户、查看资金余额、资金使用明细，以及管理账单和发票。

6.消息和通知：用户可以接收和查看系统消息，以及工单状态更新。

7.用户信息和安全：用户可以维护个人信息，如修改密码、绑定手机号和邮箱，保障账户安全。

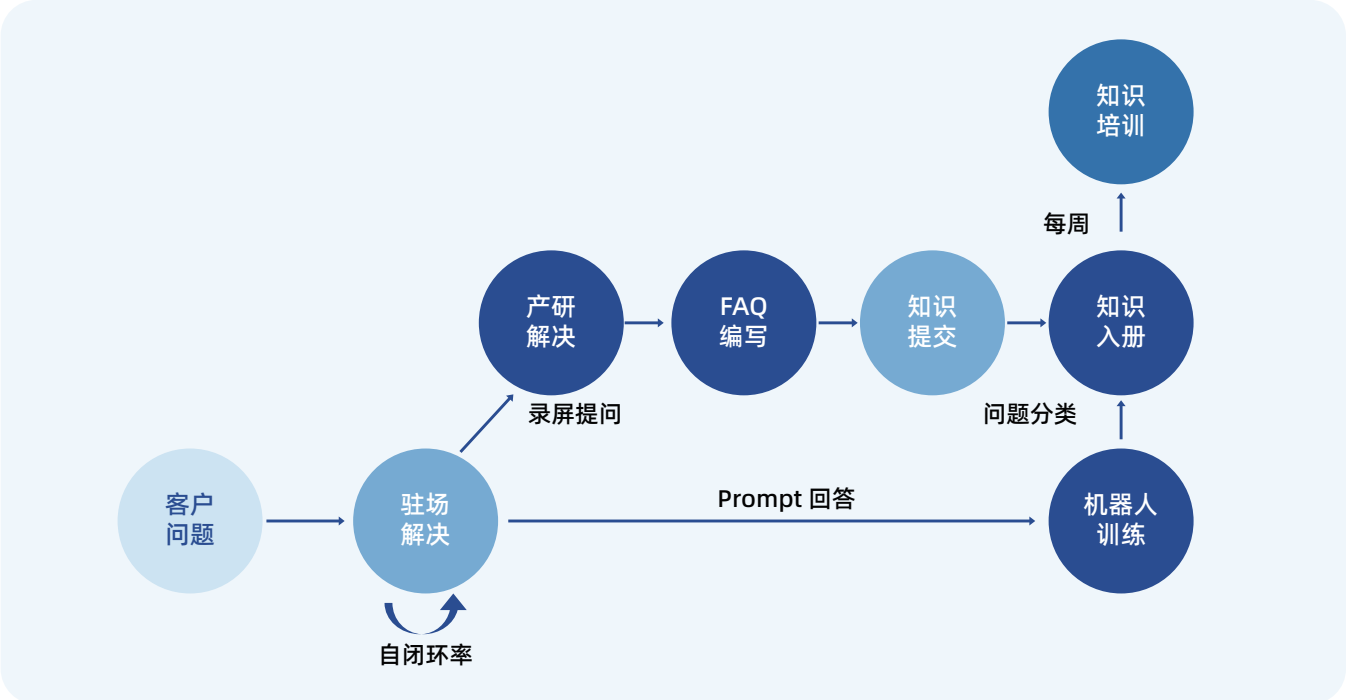
2.工单答疑

智算的工单答疑是运营难度最大的模块，用户的问题遍及平台运营、AI 模型、基础设施和上层应用。理解数据模型的用户可能看不懂网络硬件，传统运维能力难以满足 TensorFlow、PyTorch 等框架的使用问题解答。智算平台运营是云计算、大数据、人工智能平台运营的结合，要求运维人员在回答和理解客户各种的问题的时深刻理解 Linux 底层操作系统、K8s、深度学习以及镜像等专业内容。

工单答疑是影响平台客户满意度的重要服务模块，可采取根据用户画像分群体运营的模式，例如群运营、VIP 服务运营等。工单答疑需要通过对工单分类、对用户意图进行统计，从而对常见的问题及解决方案文档进行总结，通过训练自主问答机器人将结构化的正确答案输出。

3.知识库编写

知识库主包含各个产品的研发手册和使用说明、用户使用手册以及常见问题 FAQ。知识库由专门的知识管理人员做统一编写，确保知识库的质量。知识库编写完成后也需要开展知识培训给对应的用户和工作人员。



▲ 图 14 运维运营 FAQ 知识库生产流程

2.4.2 资源运营

智算运营平台根据不同类型任务对算力资源需求，提供算力纳管能力。通过建设一个高效、稳定、可靠的智算平台，为用户提供高质量的算力服务。资源运营在支持计算需求和提高资源利用率方面发挥着关键作用。

	任务类型	相关资源考虑
单机 单卡任务	适用于小型到中型的计算任务，如传统机器学习，数据加工、数据分析、图像处理等。	只涉及单卡计算资源，管理和调度相对简单，卡资源需求灵活，但容易造成整台机器的碎片化，影响算力的供给。
单机 多卡任务	适用于更高计算能力的任务，例如深度学习模型的复杂训练。	单机多卡任务可以显著提高算力，需要有效的资源规划，考虑卡内通信效率。
多机 多卡任务	适用于超大型计算任务，如超大规模深度学习模型训练、科学计算。	多机多卡任务需要复杂的集群管理和网络通信策略，确保不同节点的有效协同。

▲ 表 2 平台主要运行的任务类型和相关资源考虑

1.资源纳管

算力运营平台实现了对多样化计算资源的全面纳管，包括多种型号的GPU和CPU和定制化的计算资源。用户可以在统一的交互界面中，轻松管理整个计算平台的服务目录，实现资源的整合与优化配置。

算力运营平台支持从算力资源申请、审批、创建、变更到回收的全生命周期管理动作。平台能够精确记录资源的申请和变更记录、资源的项目归属和资源的计费主体，提供根据资源类型、作业目的、提交者身份等不同维度的资源审批能力，实现对资源的全生命周期运营管理。

2.算力调度

算力调度是指在系统中合理分配和利用计算资源的过程，其主要目的是提高整个集群的利用率，保证任务的高效执行。算力调度系统的复杂性主要由两个因素造成：一是业务资源约束因素；二是底层的基础设施、资源隔离能力约束因素。调度器的一项核心任务就是按照某一策略从集群中挑出最合适的物理机，通过机器混合调度提升机器使用效率。智算集群通过容器化的方式屏蔽了物理机之间的配置差异，进一步提升使用体验。

3.资源池化

传统的算力管理通常以物理机为单位，将物理机分配给对应团队，由相关团队内部再进行资源分配，在资源空闲时造成了极大的浪费。智算平台用虚拟化、负载均衡等技术将计算资源（如CPU、GPU、内存等）集中管理，形成一个统一的资源池，可根据资源余量、用户需求进行动态分配，提供更好的可扩展性和灵活性。同时支持队列管理能力，在资源不足的情况下开启计算任务排队模式，在有资源空闲时自动启动新任务，极大提升了资源利用和流转效率。

4.资源治理

算力的资源治理包含自定义治理策略、全链路资源治理、资源效能等方面。

自定义治理策略

根据采集的指标，结合智算应用场景，搭配贴合实际治理场景的治理策略，更精细、更精准的发现可优化的实例，治理的指标如下表所示：

	现象	原因
资源配置问题	GPU 内存/CPU 占满，但 GPU 卡未被占用	机器的内存和 CPU 被其他任务占满，GPU 实例无法启动
	GPU 卡被占用，但是 GPU 利用率为0	实例不需要 GPU 卡，建议申请低配机器
	在多卡训练任务中，长期有卡闲置	GPU 卡数申请过多，建议释放部分 GPU 卡或者更新代码
用户管理问题	空间长时间无人登陆	项目组无活跃用户
	用户欠费	用户资金不足
任务管理问题	AI 任务实例运行时间过长	运行时间久，无人管理

▲ 表 3 治理指标分类

全链路资源治理

全链路资源治理包括对治理项目的持续监控、智能推送治理建议、详细查看治理记录、实时线上反馈以及持续的校验与巡查等关键环节。运营服务团队能够通过这一机制，获得对治理状态的洞察分析，从而确保治理措施的高效和精确执行。此外，团队成员可以通过任务分配、即时在线反馈、定期巡查以及策略调整等手段，不断推动治理规则的持续运作与优化，形成良性的闭环资源治理能力。

资源效能

资源效能主要是对资源使用情况进行监控，为资源优化和管理提供数据基础和依据，并且开展对应的资源分析。和传统的机器监测重点不同，智算平台重点监测显卡性能指标。在 AI 小模型时代，由通信问题造成的性能瓶颈较为少见，而在 TB 级大模型时代，分布式训练及大规模数据可能会导致训练中断、梯度爆炸、算法重跑等问题，造成时间和成本的损失，因此资源效能模块对任务稳定性非常重要。资源效能治理主要包含以下能力：

1) GPU 性能监控：

· 实时监控显卡性能指标，包括GPU使用率、显存使用情况、温度等，以预防过热和故障。

2) 任务管理：

· 盘点当前运行的任务数量，优化任务队列，减少作业等待时间。

3) 存储监控：

· 监控系统内存和存储的使用情况，确保数据读写不会成为限制因素。

4) 网络通信：

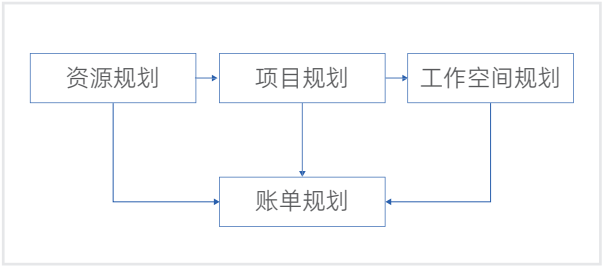
· 分布式训练中的节点间通信是关键，需要监控网络带宽和延迟。

2.4.3 运营管理

运营管理为平台在进行用户管理和资源管理时提供管理者视角，同时也助力平台从 GPU 机器的持有方转为对公众租赁算力的持续利润的 IT 运营中心。运营管理能够更加有效地处理和分析数据信息，给企业组织的决策系统提供信息支持，从而提高对平台整体的运营效率，降低额外的维护和管理成本。运营管理分为运营流程规划、数据驱动的精细化运营以及计量计费三个部分。

1.运营流程规划

运营管理人员需要依赖目前的组织现状和产品形态对计算资源、工作空间、项目组和账单等进行规划，同时也需要为用户使用平台的账号申请、资源申请/升级、工单提交、AI 资产管理、AI 任务提交、缴费、项目失效等提供流程规划和服务支持。



▲ 图 15 运营规划示意图

▲ 图 16 运营流程示意图

2.数据驱动的精细化运营

智算平台需要沉淀用户使用行为与资源运行数据，通过深度分析和挖掘，了解智算平台的运营情况及用户需求，来进行决策和优化，使运营管理团队能够更加精准地了解自身运营状况，及时调整运营策略，提升平台的使用效率。其中可以接入的数据主要包括：

1.平台内用户任务数据，例如用户在智算平台上的日活、任务失败数、工单个数。

2.机器状况，GPU 机器使用率和网络带宽。

3.自动化周报和月报和平台经营分析状况。

4.知识库文章数量，以及知识库浏览和下载量。

5.服务人员能力指标，比如变更次数、缺陷需求数、回答工单个数。

3.计量计费

智算平台计量计费需要提供、账单、对账、发票、代金券、支付、价格管理等模块。

1.订单管理：

支持生成服务订购的订单信息，订单记录服务信息、支付信息、服务开通状态。

2.账单管理：

支持月度账单的统计展示，包含消费汇总金额信息和云产品汇总金额信息。

3.发票管理：

运营方对用户提出的发票申请、撤回等操作进行审核审批以及统计分析。

4.代金券管理：

针对代金券的管理，包含限定适用的产品、用户及订单金额、支持复制代金券信息、控制代金券有效性、快速创建代金券、发放代金券和查看代金券。

5.支付管理：

支付管理支持按不同支付渠道统计查看支付金额的整体数据，包含交易金额、交易笔数、收款和退款的统计。

6.价格管理：

支持基于基础产品价格设置产品目录价格，并且能够按照用户、云产品等维度进行产品售卖折扣设置。