浪潮信息

大模型智算软件栈

OGAI

—— V1.0

2023 年 8 月

CONTENT
目录

01

大模型：AI技术新范式

大模型技术是当前AIGC技术发展的核心驱动力。从2020年GPT-3发布以来，OpenAI等国内外的科技企业和研究机构通过零样本学习（ Zero-Shot Learning ）、提示词工程（Prompt Engineering）、指令微调（SFT）、人类反馈强化学习（RLHF）等诸多技术创新，找到了有效使用大模型的技术范式。2022年底发布的ChatGPT成功引爆了公众对于生成式人工智能的热情。2023年以来，国内外针对生成式AI的投资激增，微软、谷歌等众多科技公司都在开发生成式AI模型。截止到2023年7月，国内发布的生成式AI模型已经超过了100个。

另一方面，大模型的开源开放进一步激发了学界和社区的热情。Meta在2023年3月开源的LLaMA（羊驼）大模型在短短的几个月时间内就演化出了蓬勃发展的一个大模型社区，基于LLaMA进行衍生开发的大模型包括Alpaca、BELLE、Vicuna、Koala、Orca等。此外，Falcon、MPT等众多模型的开源进一步丰富了社区生态，促进了业界对AIGC的应用落地探索。

当前，基于大模型技术的创新应用已经开始广泛地进入到日常生活和办公之中，这些系统包括大型语言模型聊天机器人，如ChatGPT和Bard，也包括办公助手MS office copilot，也包括笔记AI助手notion AI和编程助手GITHUB copilot等。

02

大模型研发与应用的需求和挑战

大模型的应用落地面临诸多挑战，而其核心是不断提高模型本身的认知、泛化、逻辑思维等各方面的基础能力，从而提高AIGC应用的智能化水平。大模型能力的提升和其训练投入的算力当量（PD，PetaFlop/s-day）正相关。根据公开资料分析，GPT-4、PaLM-2等基础模型的算力当量已经达到了GPT-3的数十倍，相当于上万颗业界性能领先的NVIDIA Hopper架构的GPU芯片组成的AI集群训练超过1个月的时间。对规模庞大的算力基础设施的需求成为了大模型研发的最大挑战。

算力平台的构建不仅仅是服务器、存储、网络等硬件设备的集成，也有诸多设备软硬件兼容性和性能调教上的know-how。需要考虑不同硬件和软件之间的兼容性和版本选择，确保驱动和工具的适配性和稳定性。比如在InfiniBand、RoCE 网络的配置和驱动安装上会遇到一些复杂的网络设置和驱动安装问题。由于涉及到用户管理，GPU 运行基础环境，并行文件系统等多个组件的安装和配置，往往需要依赖丰富的经验，整个部署过程会比较复杂。在实际的生产环境中，安装和配置集群需要兼顾性能和稳定性的考虑，为了确保系统的高性能和稳定运行，需要验证在不同的硬件环境下的软件适配，优化包括BIOS，操作系统，底层驱动，文件系统和网络等多个指标，找到最优的选择这个过程耗时耗力，容易贻误算力的上线时间。

大模型训练过程比传统的分布式训练复杂，训练周期长达数月。集群计算效力低、故障频发且处理复杂，会导致训练中断后不能及时恢复，从而会降低大模型训练的成功概率，也会使得大模型训练成本居高不下。因此，大模型对训练的稳定性、故障检测与训练容错提出了更高的要求。同时简化大模型分布式任务提交、实现智能与自动化的任务资源匹配和训练健壮性也是提升训练效率的重要保证。

在大模型的算法开发层面，从PB级数据的爬取、清洗、过滤和质检，到大规模预训练的算法设计、性能优化和失效管理；从指令微调数据集的设计到人类反馈强化学习训练的优化，冗长的开发链条意味着诸多的工程化工具的支撑。因此，如何加速模型生产、促进生成式AI落地应用，也当前业界关注的重点。

在大模型的部署与应用层面，在当前商业模型与开源模型能力表现各有专长的现状下，如何选择最为合适的基础模型，以及如何基于基础模型和行业特点，打造应用，实现大模型的落地依然是当前大模型在部署和应用上最大挑战。

03

浪潮信息的探索与实践

大模型的研发和应用涉及到多个不同的环节和团队的协同，需要硬件选型、网络设计、集群调优、算力调度、数据治理、算法架构、工程优化等多维度专业技术团队、软件工具和专家经验的支撑。

浪潮信息长期致力于人工智能算力基础设施产品的研发，其中AI服务器方面，以丰富的产品和领先的性能，市场份额常年全球领先。在产品研发、客户需求、实际应用中，浪潮信息的AI团队在AI算力系统的性能调校和优化方面积累了丰富的经验。这些经验不仅沉淀在产品上，助力浪潮信息AI服务器多年来在全球最具影响力的AI基准性能评测MLPerf的训练和推理取得了优异的成绩，也帮助客户在集群架构及软硬件层面解决了诸多如CUDA初始化失败、GPU掉卡、p2pBandwidthLatency延迟过高、NCCL通信性能低，GPU direct RDMA 未使能等问题。

浪潮信息在2021年9月发布了参数量为2457亿的中文大语言模型“源1.0”。在“源”大模型的研发过程中，浪潮信息的AI团队逐步建立了完整的从公开数据爬取到数据清洗、格式转化、数据质量评估的完整流程和工具链，并完成了5TB高质量中文数据集的清洗工作。“源”大模型的数据集和清洗经验和帮助国内不少AI团队提升了其大模型的性能表现。

在“源”大模型的研发过程中，如何高效调度千卡规模的算力，以及保障训练任务的长期稳定运行是团队重点关注和解决的一个问题。通过对云原生的调度系统进行了改造来大幅加速其启动速度，并重点解决了RDMA网络在容器中的接入和适配优化，团队较好的构建了一套能够满足大模型需求的算力调度系统。另外团队引入了多种对集群性能的监控手段和性能数据分析方法来保障训练任务的长期稳定运行。

如何提升大规模分布式训练的计算效率一直是大模型预训练的一个核心问题。特别是在实际的AI集群环境中，可能存在GPU之间的互联带宽受限或者AI服务器之间的网络互联带宽有限的情况下。 基于“源”大模型的研发经验，2022年以来，浪潮信息的AI团队协助多个客户把大模型训练的GPU峰值效率从30%左右提升到50%。从而大幅加速了模型训练过程。

04

OGAI: 大模型智算软件栈

为了满足大模型开发和应用在算力基础设施上的需求，浪潮信息发布了大模型智算软件栈OGAI“元脑生智”。OGAI（Open GenAI Infra）是浪潮信息面向以大模型为核心技术的生成式AI开发与应用场景，提供从集群系统环境部署到算力调度保障和大模型开发管理的全栈全流程的软件，从而降低大模型算力系统的使用门槛、优化大模型的研发效率，保障大模型的生产与应用。

在设计理念上，OGAI秉承全栈全流程、算力充分释放、实战验证提炼的设计原则。OGAI从当前大模型算力建设、模型开发和应用落地的实际需求出发，提供从集群环境搭建到算力调度、大模型开发的全栈软件；并覆盖大模型从数据处理到预训练和微调到多模型管理的整个研发流程。为了满足大模型计算对算力的需求，OGAI在不同的层次强调了性能优化，从服务器BIOS的调教到大规模集群组网性能和算力调度策略的多尺度、多层次的性能优化，来充分释放AI集群性能。另外，OGAI也融合了浪潮信息在MLPerf性能评测、服务客户实际需求、源大模型开发中的最佳实践。

4.1 整体架构

OGAI软件栈由5层架构组成，从L0到L4分别对应于基础设施层的智算中心OS产品、系统环境层的PODsys产品、调度平台层的AIStation产品、模型工具层的YLink产品和多模纳管层的MModel产品。

L0层智算中心OS的定位是面向智算中心等公共算力服务平台，面像多租户场景，提供灵活多样的以裸金属为主的AI算力服务。

L1层PODsys聚焦于AI集群部署场景，提供了包括基础设施环境安装、环境部署、用户管理、系统监控和资源调度一整套工具链，旨在打造一个易用、高效、开放、兼容的智算集群系统环境部署方案。

L2层AIStation聚焦于AI开发场景，通过云原生技术对集群系统中的计算资源、存储资源和网络资源进行统一的接入和纳管，提供了易于使用的开发环境和作业管理界面，并基于内置算力调度系统和训练稳定保障系统来实现易于接入、按需分配、弹性扩展和和高效稳定的AI研发应用支撑平台。

L3层YLink聚焦于大模型的开发过程，通过集成整合浪潮信息在大模型研发过程中的工具和开源工具，为用户提供高效、便捷与标准化的大模型开发与优化流程。

L4层MModel定位于多模型管理与服务平台，帮助客户更好的管理和评估模型，加速模型的部署和应用。

4.2 智算中心OS

智算中心OS是浪潮信息面向提供公共算力租赁服务的智算中心场景，满足以裸金属为主的多样化、弹性的AI算力需求为核心的智能算力运管平台。

性能优异、按需取用、灵活扩展的智能算力是大模型研发的关键，智算中心应运而生。智算中心就是要满足的不同行业、不同领域对大模型研发对算力的使用需求。不同行业、不同算力使用角色对算力的需求形式也是多种多样，如能满足性能需求的裸机算力服务。对于众多组织和角色也需要的统一管理，并做到租户隔离和数据安全性的要求。算力的规模化与多样化、用户和组织的复杂化需要一体化智算中心的运营运维平台来提供专业、高效的智算服务。

生成类AI训练业务对算力、通信性能的高要求，就要整套算力系统提供性能优越、无损耗的裸机服务，能够快速为算力用户提供标准算力服务输出。同时，大模型训练是人员参与众多、流程复杂的系统性工程，需要对多租户、多种人员角色进行协同统一的管理，对资源灵活、快速的分配。因此需要提供裸金属服务，直接为用户提供独占式的物理服务器资源，满足用户特定需求的可行性及高效性。

目前业界共识，大模型效果随着数据质量提高带来巨大提升，对高质量数据的使用和安全管理也是智算中心需要着重考虑的问题，独占式的裸金属资源能够快速解决用户极其关注的数据使用和安全问题。

智算中心OS基于通用、规范的拓展接口，以算力平台提供数据安全保护、网络安全防护和深度结合，依托智算中心 OS 的管理调度平台，实现物理计算资源（CPU、 GPU、内存、存储等）的统一管理与监控，通过智算中心OS提供生成式AI所需的无损裸机算力环境，具有计算零损耗，最大化CPU、GPU算力和高性能裸金属网络。具备批量快速部署系统能力，可在30分钟内完成百台裸金属自动配置IP、存储等。支持系统盘、数据盘使用远端存储和VPC网络，并提供远程控制台方便操作，实现基础资源即服务（IaaS）。

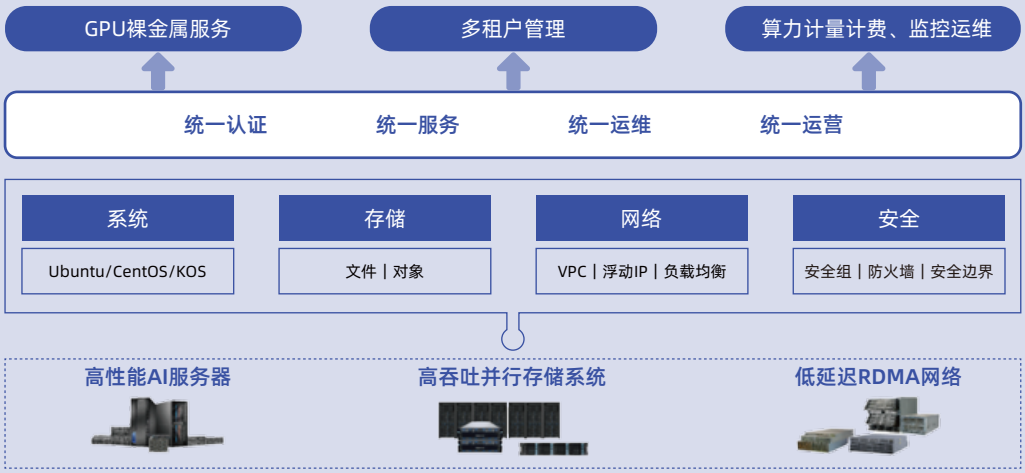
智算中心OS通过统一门户提供GPU裸金属算力服务，并提供多租户的统一门户管理，实现算力服务快速交付，在精细化的配额管理下，实现租户级别的算力配额管理，满足定制化的算力供需，并提供租户隔离、数据安全等功能。多元算力统一管理调度，避免重复建设，简化各IT系统的运维复杂度。

智算中心OS提供多样算力服务和功能形式，通过专业的计费引擎，支持资源的细粒度计费与多种计费模式。

主要提供功能如下：

- **统一门户：**通过统一门户提供 GPU 裸金属等大模型训练算力；
- **多级管理：**多租户的统一门户管理，算力服务快速交付；
- **配额管理：**细粒度的租户算力配额管理，定制化算力供需；

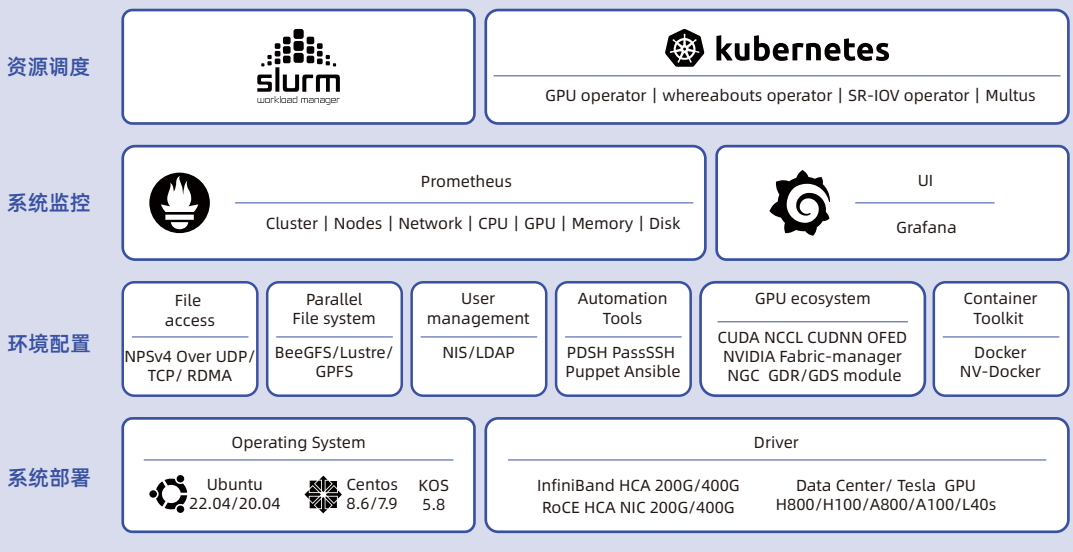
- **对外运营：**通过多种方式销售算力，提供算力服务；
- **完善健全的运维能力：**精细化监控与智能运维、全方位运维。



智算中心OS架构

4.3 PODsys

PODsys是浪潮信息聚焦于AI集群部署场景打造的一个开源项目，旨在提供一个开源、高效、兼容、易用的智算集群系统环境部署方案。PODsys提供了包括基础设施环境安装、环境部署、用户管理、系统监控和资源调度一整套工具链（https://podsys.ai/）。



PODsys 是浪潮信息聚焦于 AI 集群部署场景打造的一个开源项目，旨在提供一个开源、高效、兼容、易用的智算集群系统环境部署方案。PODsys 提供了包括基础设施环境安装、环境部署、用户管理、系统监控和资源调度一整套工具链（https://podsys.ai/）。

为了实现这个目标，PODsys整合了AI集群部署所需的数十个驱动、软件等安装包以及对应的依赖和兼容关系，并提供了一系列的简化部署的脚本工具。通过使用这些工具，用户只需执行两条简单的命令，就能完成整个集群的部署工作。

在软件包的选择上，PODsys大量选用了业界广泛使用的主流开源系统、工具、框架和软件，来保障整个部署方案的开放性和兼容性。与此同时，PODsys也基于浪潮信息AI团队的长期工作实践，挑选了最稳定和广泛兼容的软件版本，并解决了部分开源组件不可用、不兼容、不好用问题。

PODsys的主要功能包括系统部署、环境配置、系统监控和资源调度。

• **系统部署：**PODsys提供了快捷的系统部署和管理工具，包括快速安装、配置和更新集群环境。包括操作系统、NVIDIA驱动程序、InfiniBand驱动程序等必要的软件基础包，为用户提供一个完整的GPU集群环境。用户可以通过简单的命令来管理集群节点，添加或删除节点，以及监控节点的状态和性能。

•**环境配置：**PODsys支持一键部署大模型训练环境，集成了包括高速文件系统接口，自动化运维工具，NVDIA CUDA编程框架、NCCL高性能通信库，支持NGC 加速平台等功能。除此之外，PODsys具有完善的用户管理和权限控制机制。管理员可以创建和管理用户账号，分配不同的权限和资源配额。这样可以确保每个用户或团队在集群中灵活分配所需资源，并确保集群的安全性。

•**系统监控和性能优化：**PODsys提供了全面的系统监控和性能优化功能，帮助用户实时监控集群的状态和性能指标。通过可视化的界面，用户可以查看集群资源的使用情况、作业的执行情况和性能瓶颈，从而及时调整集群配置和优化作业性能。

•**资源调度和作业管理：**PODsys提供了高效的资源调度和作业管理功能，可以根据用户的需求自动调度和管理作业，确保集群的资源利用率和作业的执行效率。

针对商业用户，除了提供基础的部署和配置工具，浪潮信息也可以按需提供专家服务，协助用户基于PODsys部署集群系统环境。此外，浪潮信息的服务专家也可以协助客户完成更深入和细粒度的优化工作。这些服务涵盖了BIOS设置、操作系统优化、驱动程序调整、并行文件系统配置以及针对AI场景应用程序的优化等方面。这样，用户可以专注于模型训练和应用开发，而无需花费过多时间和精力在环境配置和优化上。

4.4 AIStation

AIStation是浪潮信息研发的商业化AI算力调度平台。AIStation面向AI开发场景，通过云原生技术对集群系统中的计算资源、存储资源和网络资源进行统一的接入和纳管，提供了易于使用的开发环境和作业管理界面，并基于内置算力调度系统来实现易于接入、按需分配、弹性扩展和和高效稳定的AI研发应用支撑平台。

在资源纳管方面，AIStation能够支持业界主流的AI加速芯片，Lustre、BeeGFS、GPFS等并行文件系统和Infiniband、RoCE等高性能网络，支持Spine-Leaf等业界常用的集群网络拓扑和管理网、计算网、存储网三网分离的集群组网架构。AIStation能够支持单一集群超千个AI计算节点、超万个AI加速芯片的接入和调度。AIStation内置数据缓存系统，大幅减少了对并行文件系统的存储IO要求，提高了训练数据读取效率，最多可以实现200~300%的性能提升。

在开发环境和作业管理方面， AIStation能够自动化配置计算、存储、网络环境，同时对一些基本的超参数提供自定义修改，方便用户使用，通过几步就能启动大模型分布式训练，目前支持诸多大模型训练框架和开源方案，如Megatron-LM、DeepSpeed等。帮助开发者在大规模集群环境下便捷地提交分布式任务，并针对分布式任务对GPU算力的实际需求，以存储亲和、设备亲和、网络拓扑亲和等多种亲和性调度策略，降低了构建分布式训练任务的技术门槛。通过构建专用调度模式、优化训练任务运行模式，让AI从业者更便捷地使用大规模算力，运行大模型的开发与训练任务。

在大规模算力调度方面，AIStation算力调度器通过动态、智能地管理和调配集群计算资源，制定合理的作业执行计划，以最大限度地利用资源，满足训练任务的时延和吞吐需求。AIStation优化调度系统性能，实现了上千POD极速启动和环境就绪。针对大模型训练通信要求高的场景，AIStation提供集群拓扑感知能力，容器网络与集群物理网络一致，保证了容器互联性能，满足训练通信要求。AIStation支持交换机级别的资源调度，减少跨交换机流量，保障了大规模集群的网络感知和有效调度，避免网络拥堵问题，提高通信效率。通过对分布式通信算法和集群网络拓扑的联合优化，AIStation在千卡规模集群中能实现90%以上的分布式计算扩展比。尤其AIStation对大规模RoCE无损网络下的大模型训练也做了相应优化，实测网络性能稳定性达到了业界较高水平。

在训练稳定保障方面，健壮性与稳定性是高效完成大模型训练的必要条件。针对大规模分布式计算，AIStation通过内置的监控全面的监控系统和智能运维模块，能够快速定位芯片、网卡、通讯设备异常或故障。针对正在训练大模型任务进行暂停保持，然后从热备算力中进行自动弹性替换异常节点，最后健康节点进行快速checkpoint读取，实现断点自动续训。

AIStaiton的训练全生命周期监管能够实现计算集群、大模型训练异常的全自动化处理。同时满足大模型训练的诸多诉求，如资源使用视图、计算与网络调度策略、分布式训练加速、训练监控、训练断点自动恢复能力，保证了训练的稳定性和效率。

4.5 YLink

YLink聚焦于大模型的数据治理、预训练、微调等开发过程，通过集成整合浪潮信息在大模型研发过程中的工具和开源工具，为用户提供高效、便捷与标准化的大模型开发与优化的工具和流程。

YLink聚焦于大模型的数据治理、预训练、微调等开发过程，通过集成整合浪潮信息在大模型研发过程中的工具和开源工具，为用户提供高效、便捷与标准化的大模型开发与优化的工具和流程。

大模型开发过程中的2个核心作业环节就是数据处理和模型训练。其中数据处理将通过对数据源和元数据进行采集、处理等操作后，转换成模型训练所需类型数据的重要步骤，也是大模型开发服务生命周期的上游环节，高质量的数据有助于生成更优质的模型。模型训练是基于数据集和算法进行训练，包括：算法选择、超参数优化、模型评估和选择等环节，最终输出已训练模型。如何使用各种并行策略提升分布式训练效率、减少训练中断频率对于保障大模型构建起着至关重要的作用。

当前，YLink集成的工具包包括数据处理工具包（Y-DataKit）、大模型训练工具包（Y-TrainKit）和大模型微调工具包（Y-FTKit）。此外，基于这些工具，YLink也提供了大模型预训练和微调的参考流程。模型预训练过程可以表示为：数据集+模型代码+训练方法（含算力、训练策略）=预训练大模型，涉及的开发工具包括：数据处理工具（数据采集、数据清洗、数据转化工具）、模型训练工具（分布式训练框架及配置脚本）。针对大模型的预训练过程，YLink具体提供了数据处理工具Gather、Transform和Purity以及基于业界主流大模型分布式训练框架 NVIDIA Megatron 和 MS DeepSpeed的大规模分布式预训练参考流程。

● 数据处理工具包（Y-DataKit）

• Gather：数据处理工具， 支持从Web、知识文档等3种安全数据源进行不间断、高速收集数据

• Transform：格式转换工具，支持对PDF、Docx、XML等8种原始格式数据及数学公式进行高准确率转换

• Purify：数据清洗工具，提供内容主题、标点、分词、符号、段落等28种数据内容清洗能力

● 大模型训练工具包（Y-TrainKit）

• NVIDIA Megatron：分布式训练框架，基于PyTorch框架，用于训练Transformer 架构的巨型语言模型

• MS DeepSpeed：分布式训练框架，开源深度学习优化库，旨在提高大规模模型训练的效率和可扩展性

模型微调过程可以表示为：微调数据+预训练模型+微调算法（算力、微调框架）=任务/领域模型，其中微调数据的类型主要是

指令微调（SFT）数据，相较于预训练数据而言，数据质量要求更高，且需要以类似问答对的形式提供给模型。微调算法包括全量微调、高效微调（PEFT）和强化学习等。针对大模型的微调过程，YLink提供了DataGen和FileQA两个SFT数据数据生成工具。以及基于这些数据生成工具和大模型微调框架LMFLow、提示词优化工具Prompter的大模型微调参考流程。

● 数据处理工具包（Y-DataKit）

• DataGen：数据生成工具，支持基于主题、指令词与大语言模型APIs调用，快速生成高质量微调（SFT）数据集

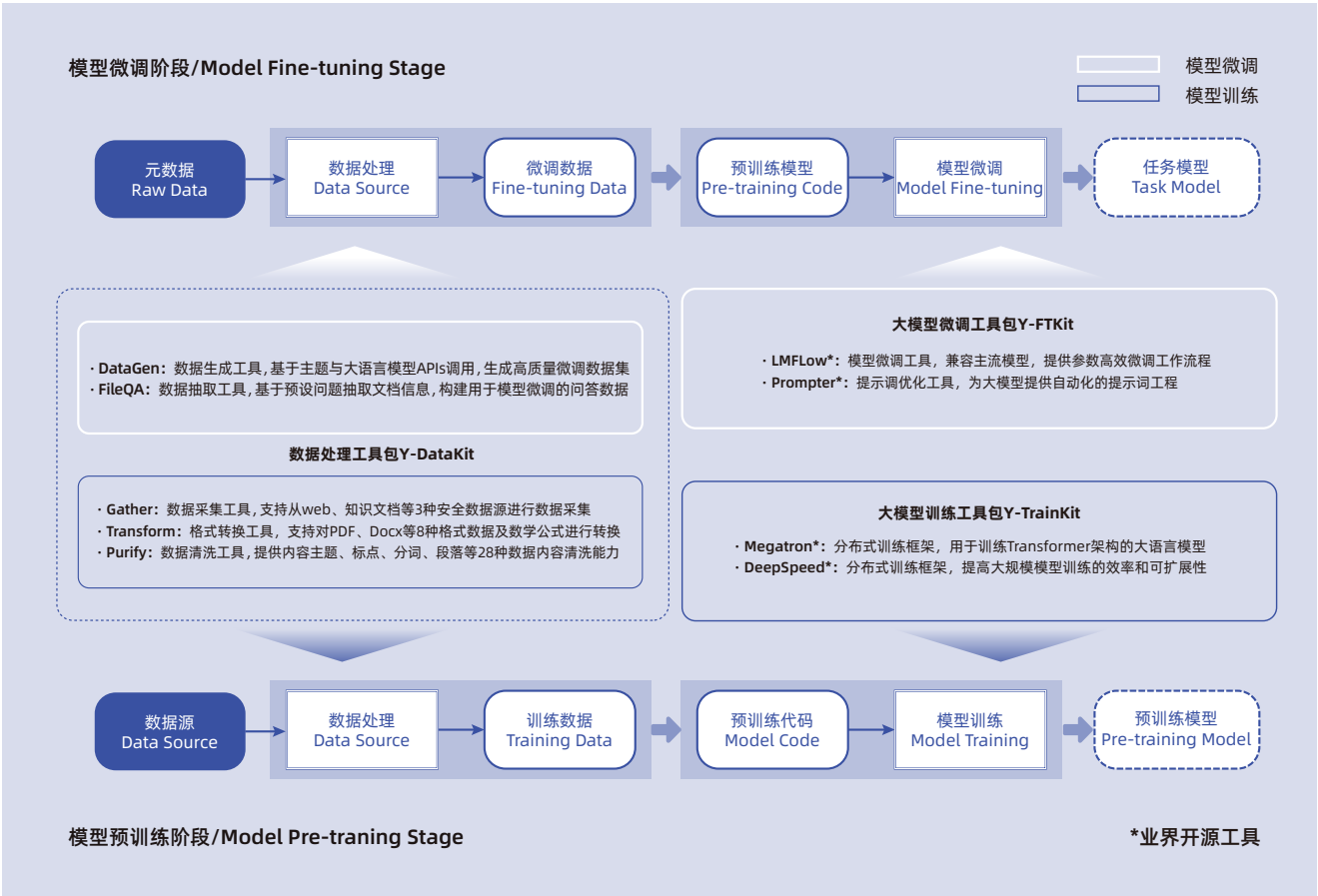
• FileQA：数据抽取工具，可以通过固定文档（PDF、Docx等）基于预设问题抽取数据（或答案），形成有效问答对形式，用于模型微调数据构建

● 大模型微调工具包（Y-FTKit）

• LMFLow：模型微调工具，提供完整的微调工作流程、指令调整、参数高效微调，支持开源模型与“源”大模型微调

• Prompter：提示词优化工具，为大语言模型提供自动化的提示工程，使大语言模型生成高质量内容

在OGAI大模型智算软件栈中，YLink作为大模型开发工具链的能力支持层，面向大模型生成的重要环节提供了标准化的数据处理工具、高效的分布式训练框架与多种开箱即用的SOTA模型微调算法，可以有效提升模型预训练效率与的微调质量。



4.6 MModel

多模型的纳管，可以帮助研究人员、开发者和从业者更好地管理多版本、多类型的基础大模型与任务模型，在保证模型权重、数据集安全的前提下提供对外APIs服务的能力，并且用户可以基于平台的多模型评测服务，使用多样化的评测数据集与评测任务同时对多个模型进行生成准确率、推理延迟、推理稳定性等指标进行全面的评估，为使用人员与应用场景提供准确、快速和较高参考价值的指导。MModel是一个为大模型开发者和研究人员提供的具备多模型接入、服务、评测等功能的纳管平台，致力于使大模型的安全管理、便捷使用与能力测评变得更加高效和易用，MModel的核心组件包括：数据集管理、模型纳管和评测。

- **数据集管理：** MModel 提供了用于微调与评测的语言、推理、知识等多分类数据集，可以快速加载和处理各种NLP数据集。并提供高效的数据处理和数据集定向下载功能，使得开发者可以更专注于如何快速、有效选取适合使用场景的大模型。

- **模型纳管：** MModel 提供了多规格、多版本、多类型的基础模型和任务模型的纳管能力，适配并预置一系列经过大规模预训练的高质量模型，这些模型在多种NLP任务上表现出色，同时提供了包括多种主流开源模型和闭源模型的APIs接入能力，使用户可以在保证模型权重安全的基础上进行模型的权限、版本管理，提高了模型使用与测试的便捷性。

- **模型测评：** 面向大模型的一站式基准评测能力，MModel提供在线基准评测，可支持语言、知识、推理、多学科等模型评测方案，并支持大模型推理性能评测，用户可通过测试报表查看大模型的推理延迟、稳定性测试和其他关键指标，也可扩展为面向多路大模型横向基准测评。

- **生态合作与资源共享：** MModel 鼓励元脑生态间的合作和资源共享。开发者可以通过平台分享自己的模型、数据集和代码，从而促进了领域内知识的传播与共享。