Enabling LLM Adoption at the Enterprise

As enterprises increasingly adopt AI and machine learning (ML) to drive business value, there is a growing need for streamlined and efficient LLM (large language model) workflows. Run:ai provides a comprehensive platform for end-to-end LLM lifecycle management, enabling enterprises to fine-tune, prompt engineer, and deploy LLM models with ease.

Single Platform for End-to-End LLM Lifecycle

Manage the entire LLM lifecycle within a single platform. Data scientists can fine-tune their models, prompt engineer their datasets, and deploy their models with ease.. This unified approach streamlines the LLM workflow, reducing the time and effort required to deploy high-quality models.

Deploy LLM with Confidence

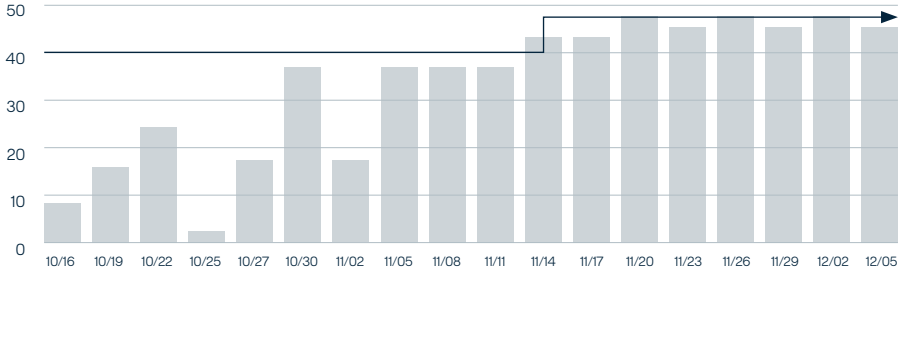
Easily run large-scale LLMs on GPUs across cloud and on-premises, that automatically scale based on SLA requirements. This ensures that LLM workloads are deployed with confidence, knowing that they have the necessary compute resources to perform optimally. Additionally, enterprises can configure guaranteed quotas to ensure that LLM workloads receive the necessary resources.

Resource Management

Pool resources in a single cluster and manage the lifecycle of multiple models in an effective (and cost effective) way.. Platform teams can set policies and prioritizations to ensure that LLM workloads are optimized for performance while balancing the needs of other workloads. This approach reduces infrastructure costs and complexity while maximizing resource utilization.

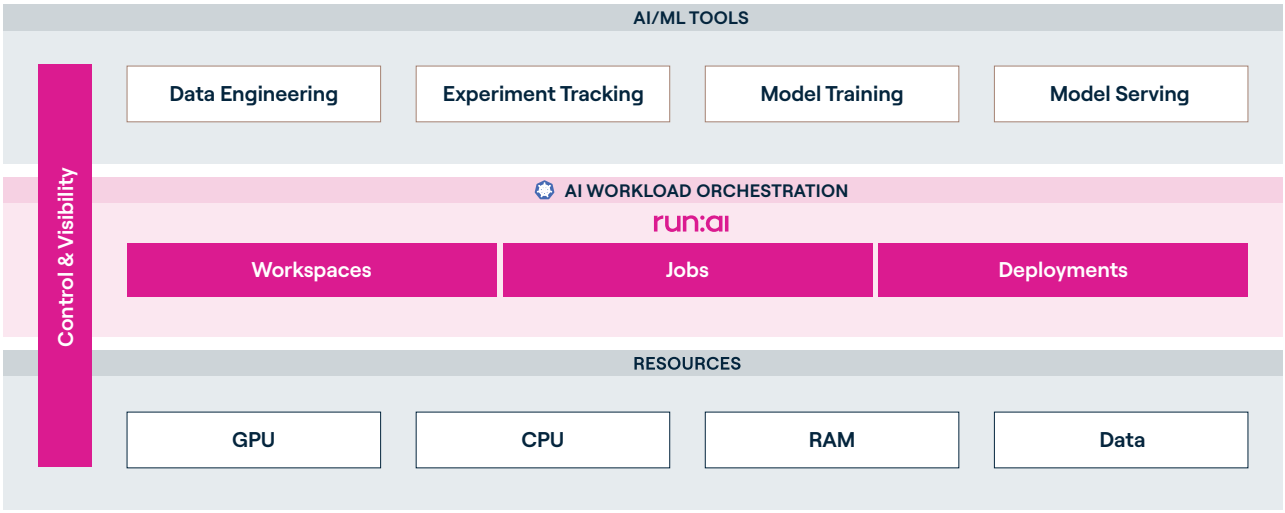
5X

UTILIZATION



Platform Overview

The Run:ai Atlas platform sits in between the infrastructure and the AI workloads that require access to these valuable resources. Platform teams gain centralized control and visibility across all AI infrastructure, whether on-premises or cloud. AI/ML teams get streamlined and self-service access to all the compute they need, when they need it, using the tools they prefer.



Feature Highlights

Control & Visibility

Gain insights with real time and historical analytics of all LLMs and their resources managed by the platform. Simplify management and enable automation by defining policies around access to LLMs and the resources they consume.

Autoscaling

Automatically scale model deployments up or down based on predefined thresholds using built-in or custom metrics, ensuring model SLAs are met and results in an optimal end-user experience.

AI Workload Scheduler

Run:ai's K8s Scheduler uses multiple queues to manage batch tasks, with customizable rules and policies for each queue based on business priorities. Combined with over-quota and fairness policies, resource allocation is automated and optimized for maximum cluster utilization.

Customers Accelerating AI with Run:ai Atlas

SONY BNY MELLON ZEBRA xiaomi