

오픈소스를 통한 AI의 대중화

성공적인 생성형 AI 애플리케이션 구축 전략

한국 레드햇
테크세일즈팀

Associate Principal Solution Architect / 상무
유혁 (hyoo@redhat.com)



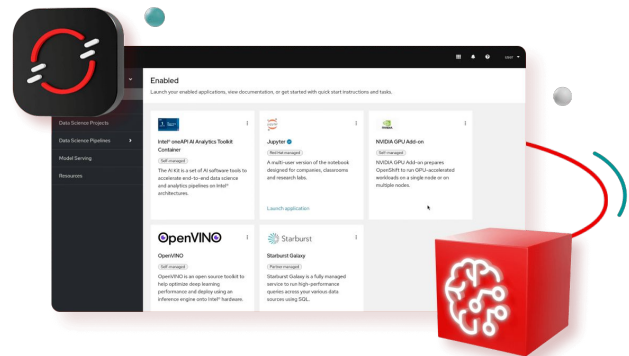
유혁 / Associate Principal Solution Architect

- 2019년 레드햇 입사 (5+ yrs at RH)
- Telco 어카운트 담당
- 레드햇 입사 전부터 각종 Telco 사업을 담당해옴:
 - VoIP, IMS, VoLTE, SDN, NFV 등
- 오픈소스를 통한 Telco 혁신에 주력

This section includes:

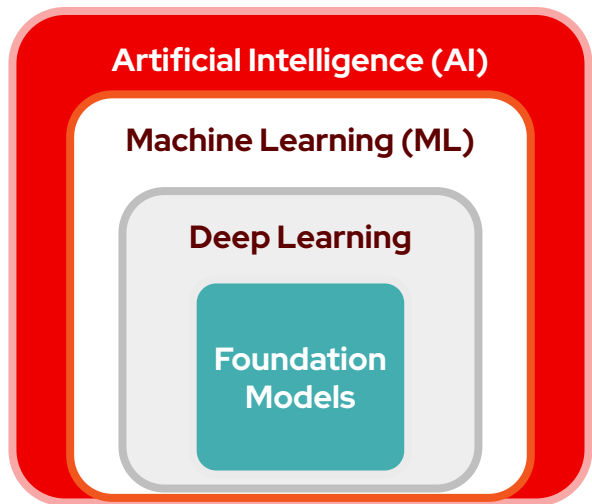
- ▶ 배경
- ▶ 레드햇의 AI 전략
- ▶ 오픈소스를 통한 AI의 대중화
- ▶ 레드햇 AI 포트폴리오

배경



AI에서 파운데이션 모델(Foundation Models)의 역할

Foundation models은 Chat GPT 같은 다양한 범위의 다운스트림 AI앱의 기반을 제공

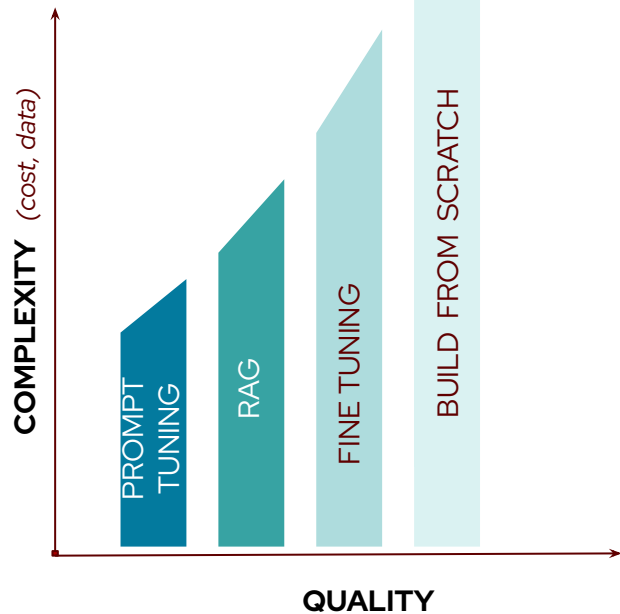


파운데이션 모델은 딥러닝 기술을 사용하여 광범위하고 다양한 데이터 세트에 대해 학습된 대규모 사전 훈련(pre-trained) 모델입니다.

- ▶ 일반적인 특징과 패턴 학습(일반 지능)
- ▶ 인간과 유사한 콘텐츠 생성(GenAI 앱)
- ▶ 다양한 AI 모델 앱의 기반이 됩니다(AI 지원)

거의 즉시 사용할 수 있는 특성으로 인해 실제 비즈니스에서 AI를 빠르게 사용할 수 있습니다

Foundation Model을 실제 이용하기 위해서는 더 많은 작업이 필요



- ▶ **Prompt tuning** allows to adapt models offering 'good enough' accuracy but doing it with less resources
- ▶ **Retrieval augmented generation (RAG)** allows training models with targeted information without modifying the underlying model itself
- ▶ **Fine tuning foundation models** requires a high amount of resources (data, hardware, people)
- ▶ **Training a Foundation Model from scratch** requires un-realistic amount of computing, and goes against the principles of re-using foundation models.

데이터에 따라 달라지는 AI 모델의 정확도

선진 기업들이 제공하는 LLM은 공개된 데이터를 기반으로 만들어지고 있다.

기업이 AI 활용에 필요한 많은 내부 데이터는 LLM에 포함되지 않는다.

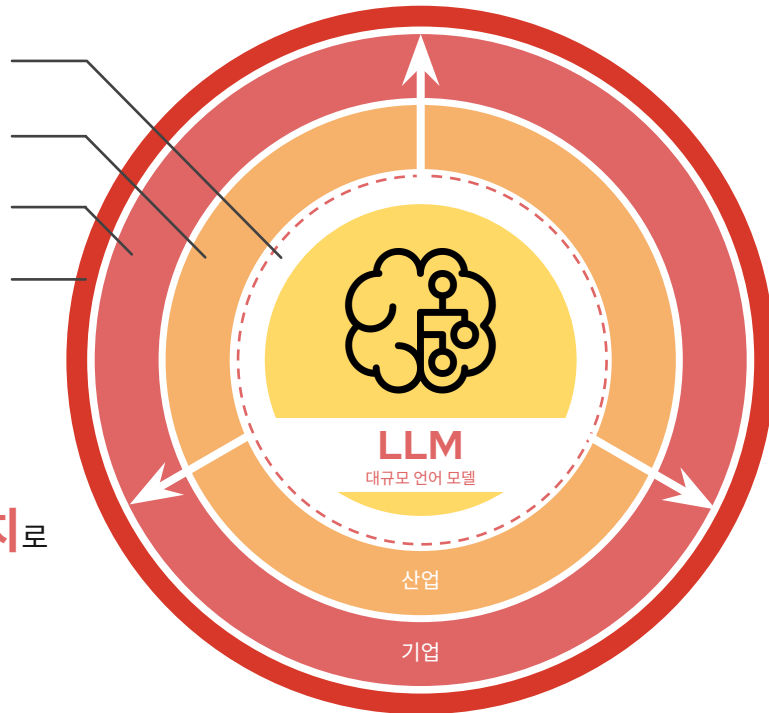
시장 데이터

산업 데이터

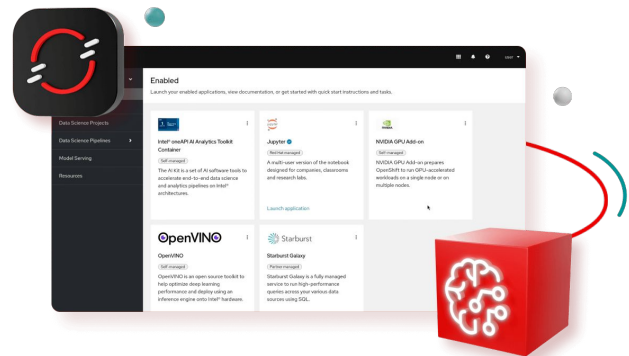
기업 데이터

개인 데이터

LLM에게 기업의 일을 맡기는 방법,
내부 데이터를 바탕으로 **자체 AI를 키우는 것이 가장 큰 가치**로 이어진다.



레드햇의 AI 전략



AI에 대한 Red Hat의 접근 방식

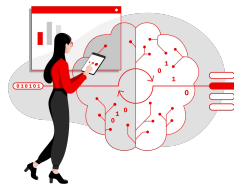
AI + 오픈 소스 커뮤니티



AI 오픈 소스 커뮤니티에 대한 Red Hat의 기여와 커뮤니티 중심의 새로운 LLM 개발 모델 제안

- 1 IBM Research와 협력하여 오픈 소스화한 Granite
- 2 InstructLab의 개방형 LLM 개발 모델

엔터프라이즈 오픈 소스 AI

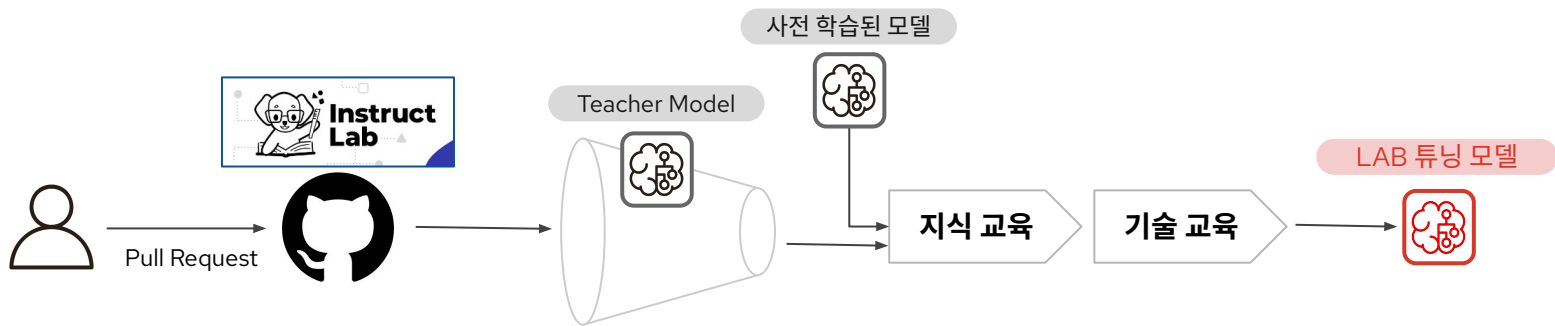


AI 혁신을 비즈니스 가치로 연결하는 Red Hat의 제품 전략

- 1 AI로 핵심 제품 강화
- 2 AI 개발 및 운영을 위한 플랫폼 제공

AI의 대중화를 목표로 하는 InstructLab

AI + 오픈 소스 커뮤니티



1단계 커뮤니티 데이터 수집

기여자는 학습의 기초로 사용될 데이터를 추가하고 GitHub의 InstructLab [리포지토리](#)에 pull request를 제출합니다.

2단계 데이터 증강/필터링

Teacher Model(LLM)은 추가된 데이터를 기반으로 데이터를 확장하고 필터링합니다.

3단계: 확장 데이터를 사용한 LLM 추가 학습

2단계에서 생성된 데이터를 사용하여 사전 학습된 LLM에 대해 두 단계의 미세 조정을 수행하여 새로운 지식과 기술을 습득한 LLM을 생성합니다.

Red Hat: 비즈니스에 AI 배포

엔터프라이즈 오픈 소스 AI

1

AI를 활용한 핵심 제품 강화

작년에 일반 사용 가능했던 앤서블 라이트스피드 메커니즘이 RHEL 및 오픈시프트에 확장되어 Red Hat 라이트스피드로 발표되었습니다.



Red Hat 라이트스피드



OpenShift **LightSpeed**



Red Hat Enterprise Linux **LightSpeed**



Ansible **LightSpeed**

생성형 AI는 제품 사용에 대한 기술적 장벽을 낮추고 개발자와 운영자의 생산성 향상에 기여합니다.

2

AI 개발 및 운영을 위한 플랫폼 제공

RHEL AI 및 Podman AI Lab과 같은 신제품과 기존 OpenShift AI의 새로운 기능 및 파트너십 확장을 발표했습니다.

Red Hat AI 플랫폼



Red Hat
Enterprise Linux AI

로컬 환경에서 AI 개발 기능 제공

- OSS Granite 제공
- InstructLab CLI
- RHEL 이미지 모드

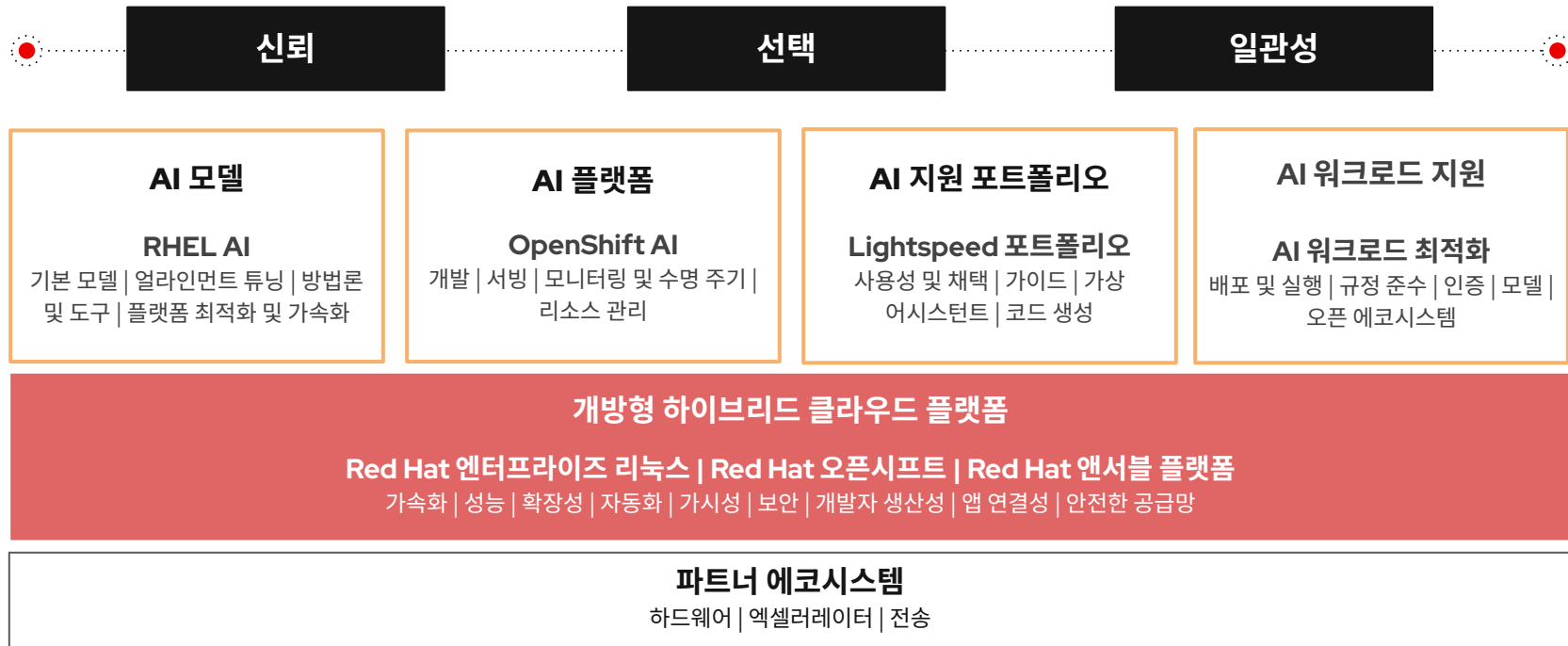


Red Hat
OpenShift AI

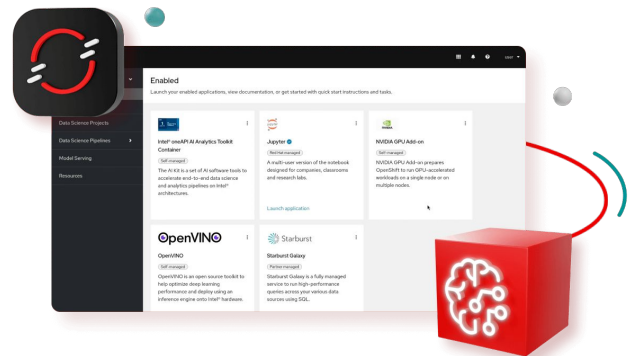
통합 MLOps 플랫폼 제공

- 엣지에 모델 배포
- 분산 교육
- vLLM 지원
- 파트너십 발표

Red Hat의 AI 포트폴리오 전략



오픈소스를 통한 AI의 대중화



AI 게임의 판도가 바뀌었습니다 - it's now free and open source

InstructLab은 자신의 지식을 활용하여 AI를 육성합니다

커뮤니티 중심의 개발

데이터에서 추론까지 오픈소스로

AI 윤리 기준 적용

상업 및 개인 용도로 무엇을 얻을 수 있는지 알아보세요



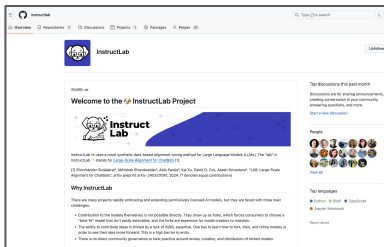
InstructLab

Granite base model

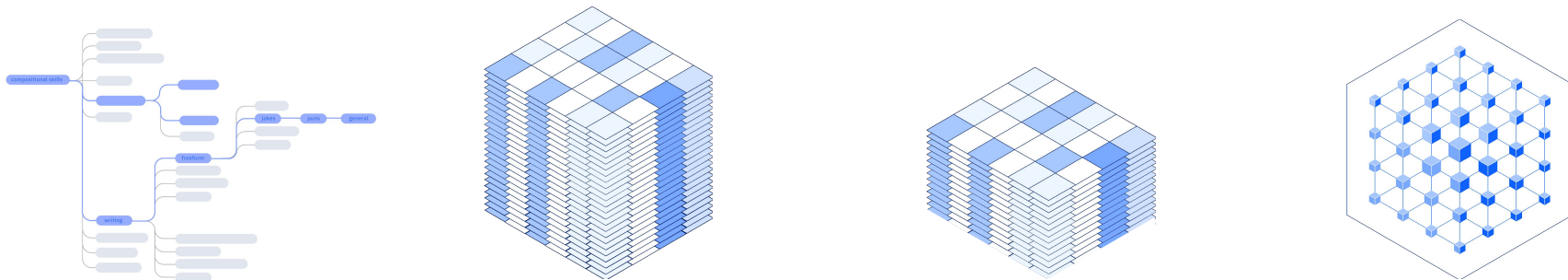
<https://huggingface.co/ibm-granite/granite-7b-base>

Apache 2.0 라이선스 하에 공개

지속적인 진화



LAB (Large-scale Alignment for ChatBots) Method



Taxonomy에 기반하여 skill & knowledge를 표현

누락된 모델 지식이나 기술을 계층적 분류법(Taxonomy)로 표현하고, 누락된 기술당 누락된 동작을 나타내는 5개 이상의 예시 데이터 포인트를 제공합니다.

Teacher model로 합성 데이터(Synthetic data) 생성

Teacher model은 분류법 전반에 걸쳐 수백만 개의 질문과 답변으로 구성된 "커리큘럼"을 생성합니다.

Critic model로 생성된 합성 데이터를 검증

Critic models은 정확성과 품질을 위해 질문을 필터링합니다. 합성 데이터에 부적절한 자료가 있는지 스캔합니다.

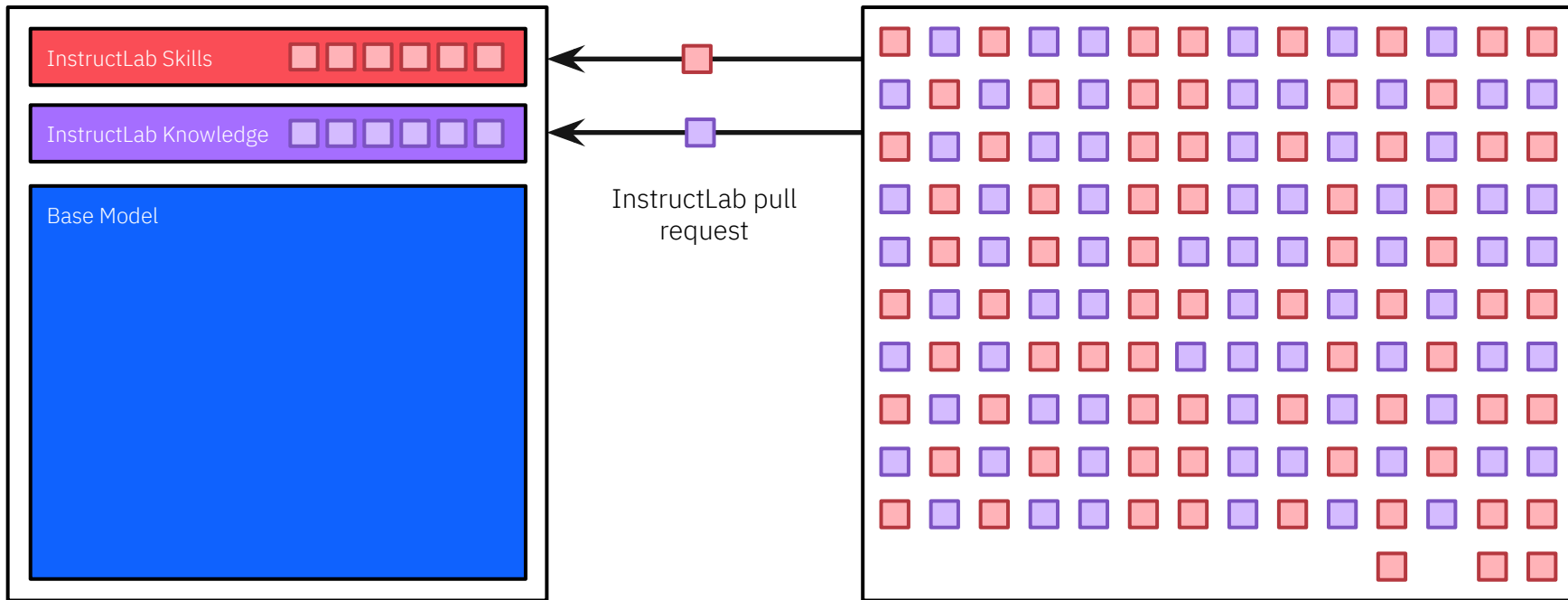
student model 상에서 skill, knowledge를 학습

Student model은 새로운 학습 (Training) 방식을 사용하여 커리큘럼을 통해 학습됩니다.

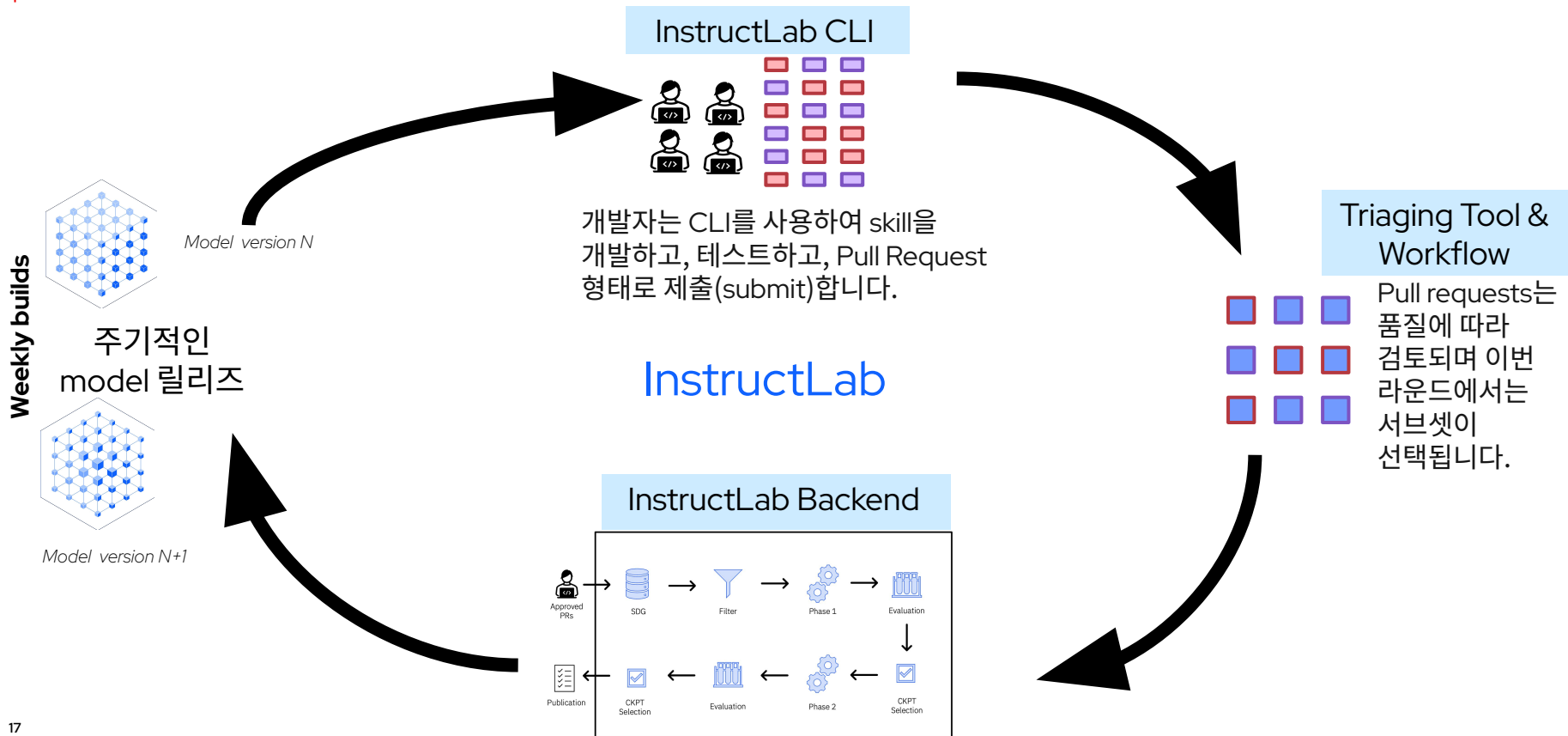
InstructLab: 커뮤니티 중심의 개발 및 모델 진화를 위해 LAB method를 활용합니다

Model 스택

커뮤니티는 skill recipe를 생성하고 기여



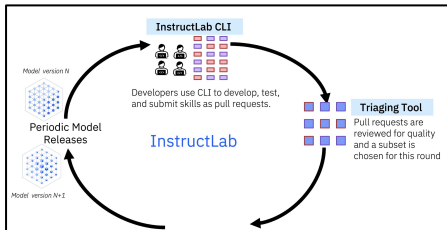
InstructLab: 신속한 오픈소스 혁신을 위한 엔진 (커뮤니티 & 기업)



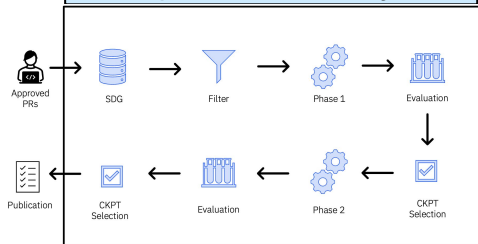
분류된(Triaged) pull request는 백엔드 플로우(합성 데이터 생성 + 다단계 학습)을 실행하는 데 사용됩니다.

InstructLab 커뮤니티 인스턴스 vs. 고객 인스턴스

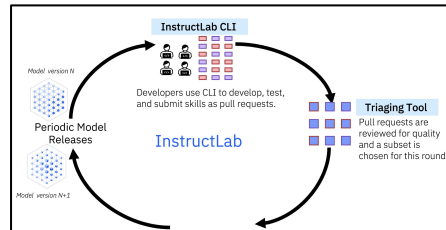
Community Instance
operated by Red Hat with support
from IBM Research team



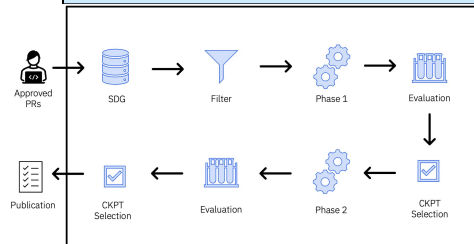
InstructLab Backend for
open community



Customer Instance
operated by Customer or RedHat/Partner,aaS
with commercial support from Red Hat



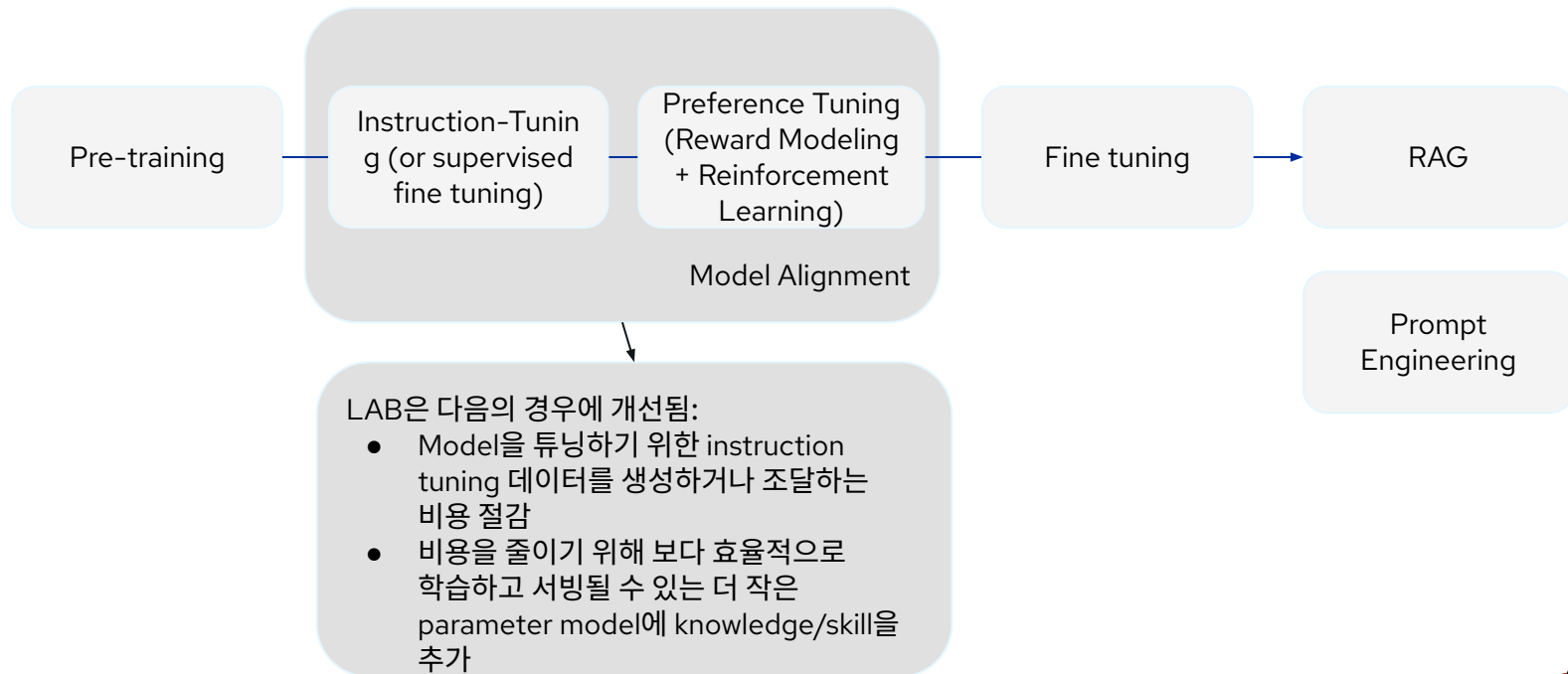
InstructLab Backend for
customers



InstructLab vs. Fine tuning

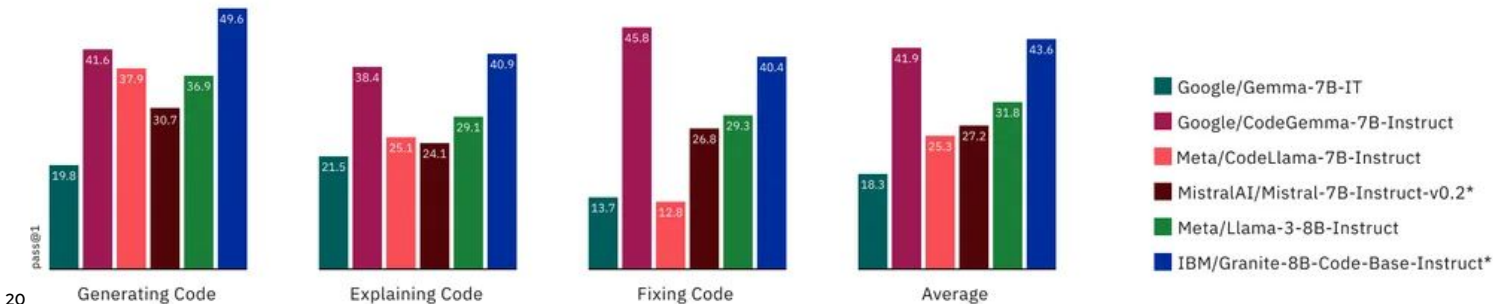
일반적으로 Model provider가 수행

일반적으로 Model consumer/user가 수행



Granite Model 성능 비교

Granite model은 Code Llama와 같은 두 배 사이즈의 모델보다 우수한 성능을 보임. 일부 타 model이 코드 생성과 같은 일부 작업에서 보다 좋은 성능을 보여줬으나 generation, fixing, explanation 측면에서 Granite 보다 높은 수준의 성능을 보여준 모델은 없었음.

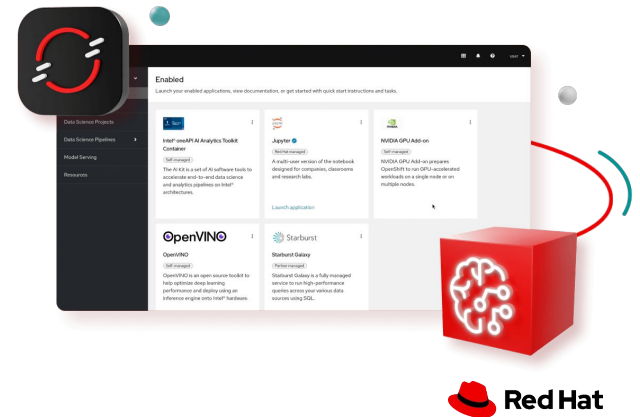


Benchmarks used:

HumanEvalPack,
HumanEvalPlus,
RepoBench,

across most major
programming languages,
including Python,
JavaScript, Java, Go,
C++, and Rust.

레드햇 AI 포트폴리오



AI 플랫폼 포트폴리오

InstructLab은 사용자가 자신의 환경에서 LLM을 쉽게 실험할 수 있는 오픈 소스 커뮤니티 프로젝트입니다. RHEL AI 및 OpenShift AI를 사용하면 단일 서버 또는 여러 서버로 구성된 클러스터에서 엔터프라이즈급 LLM 학습 및 추론이 가능합니다.



Laptop / Desktop



InstructLab

소규모 데이터 세트에 대한 제한된 데스크톱 규모 학습 방법(qlora)을 통해 학습 및 실험 가능

Podman Desktop과의 통합(To-be)

Single Server



Red Hat
Enterprise Linux AI

단일 서버에서 Teacher Model을 통한 완전한 데이터 확장 및 필터링을 통한 엔터프라이즈급 학습

전체 합성(synthetic) 데이터 생성, Teacher 및 critic 모델을 사용한 운영환경 레벨 모델 학습 기능 제공

스크립팅 가능한 기본 요소에 초점을 맞춘 도구

Cluster

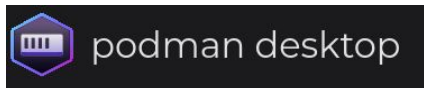


Red Hat
OpenShift AI

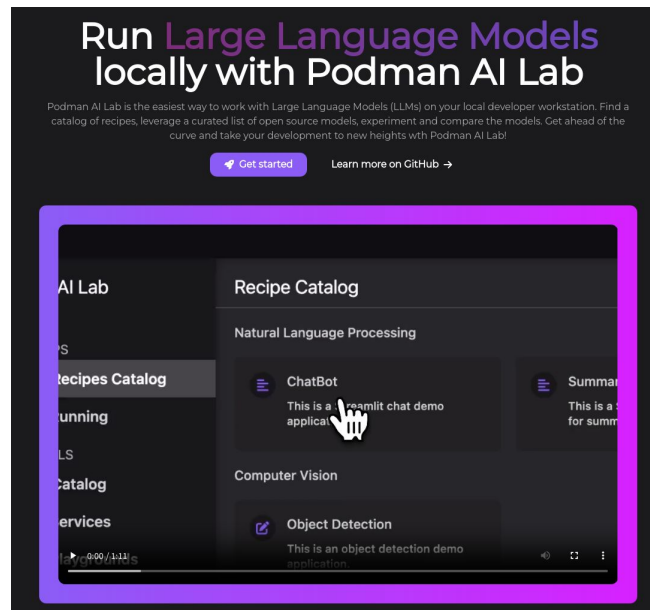
RHEL AI이 제공하는 기능을 모두 포함하며, Kubernetes 확장, 자동화 및 MLOps 서비스의 모든 기능 제공

운영환경 레벨 모델 학습 기능 제공

InstructLab + Podman Desktop AI Lab



로컬 PC에서 생성형 AI를 사용하여 애플리케이션 개발 장벽을 낮추는데 사용 가능한 Podman Desktop 확장 도구 세트를 개발하였습니다.





RHEL AI

Foundation Model Platform

Granite family LLM (Large Language Model)을
원활하게 개발, 검증 및 실행하여 엔터프라이즈
app을 발전시킵니다

The model is the new platform.



Open Granite models

고성능, 완전한 오픈 소스, 커뮤니티의 협업을 통해 개발된 Granite 언어 및 코드 모델은 Red Hat과 IBM의 완벽한 지원 및 보증



InstructLab model alignment

다양한 응용 프로그램을 위한 LLM 역량을 효율적으로 강화하고 광범위한 사용자가 지식 및 기술 기여에 접근할 수 있도록 하는 확장 가능하고 비용 효율적인 솔루션



최적화된 bootable model runtime 인스턴스

Pytorch/런타임 라이브러리, Nvidia, Intel 및 AMD를 위한 하드웨어 최적화 추론을 포함하여 어디서나 실행할 수 있고 확장성과 수명 주기를 위한 OpenShift AI로의 진입로를 제공하고 에이전트 통합 및 거버넌스를 위한 WatsonX를 제공하는 부팅 가능한 RHEL 이미지로 패키징된 Granite 모델 및 InstructLab 도구



엔터프라이즈 기술지원, 라이프사이클 및 보증

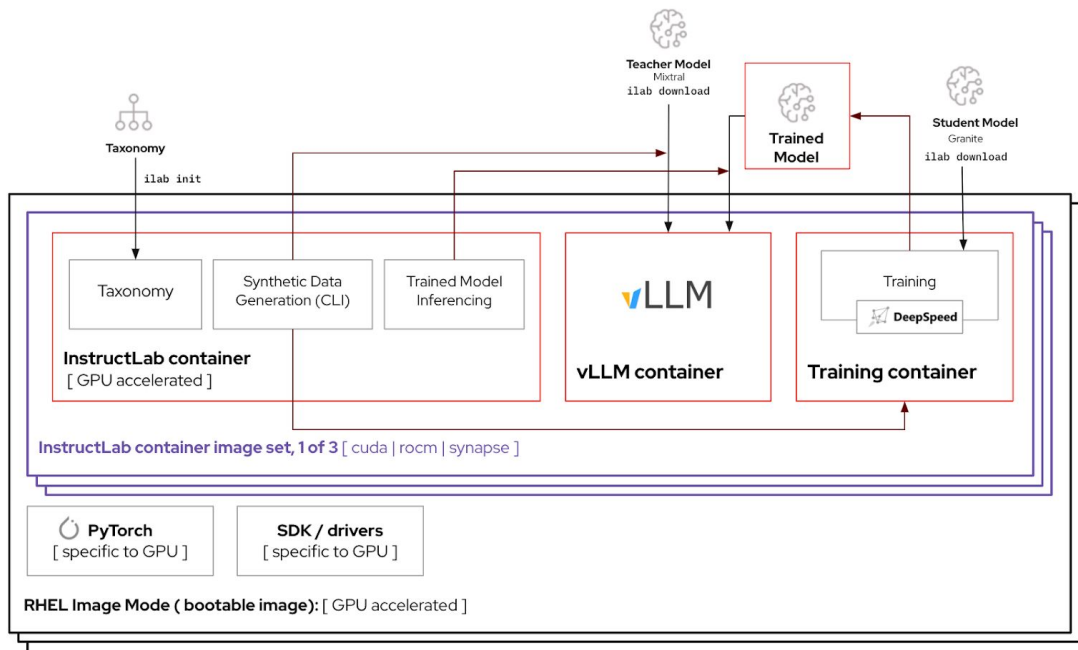
신뢰할 수 있는 엔터프라이즈 플랫폼, 24시간 연중무휴 생산 지원, 확장된 모델 라이프사이클 및 모델 IP 보증

RHEL AI 컴포넌트



Red Hat Enterprise Linux AI

- [Press Release](#)
- [RHEL AI blog](#)
- [InstructLab Community Page](#)
- [RHEL AI developer preview](#)
- [LAB: Large-Scale Alignment for ChatBots](#)
- [IBM research blog on Synthetic training for LLMs](#)



All logos and marks are the property of their respective owners. Details in the appendix.



Red Hat OpenShift AI

Hybrid MLOps platform

IT, Data Science 및 Application 개발 팀을 하나로 모으기 위해 공통 플랫폼 내에서 협업하세요

Available as

- 클라우드 서비스
- 전통적인 소프트웨어 제품 (on-site or in the cloud)



Model tuning & alignment

PyTorch와 InstructLab을 비롯한 핵심 AI/ML 라이브러리와 프레임워크에 접근하여 JupyterLab에서 탐색적 실험을 수행해 보세요. 노트북 이미지나 직접 만든 이미지를 사용할 수 있습니다.



Model serving & monitoring

모든 클라우드, 완전 관리형 (fully managed), 자체 관리형 (self-managed) OpenShift 환경에 모델을 배포하고 확장하며, 모델의 성능을 중앙에서 모니터링합니다.



라이프사이클 관리

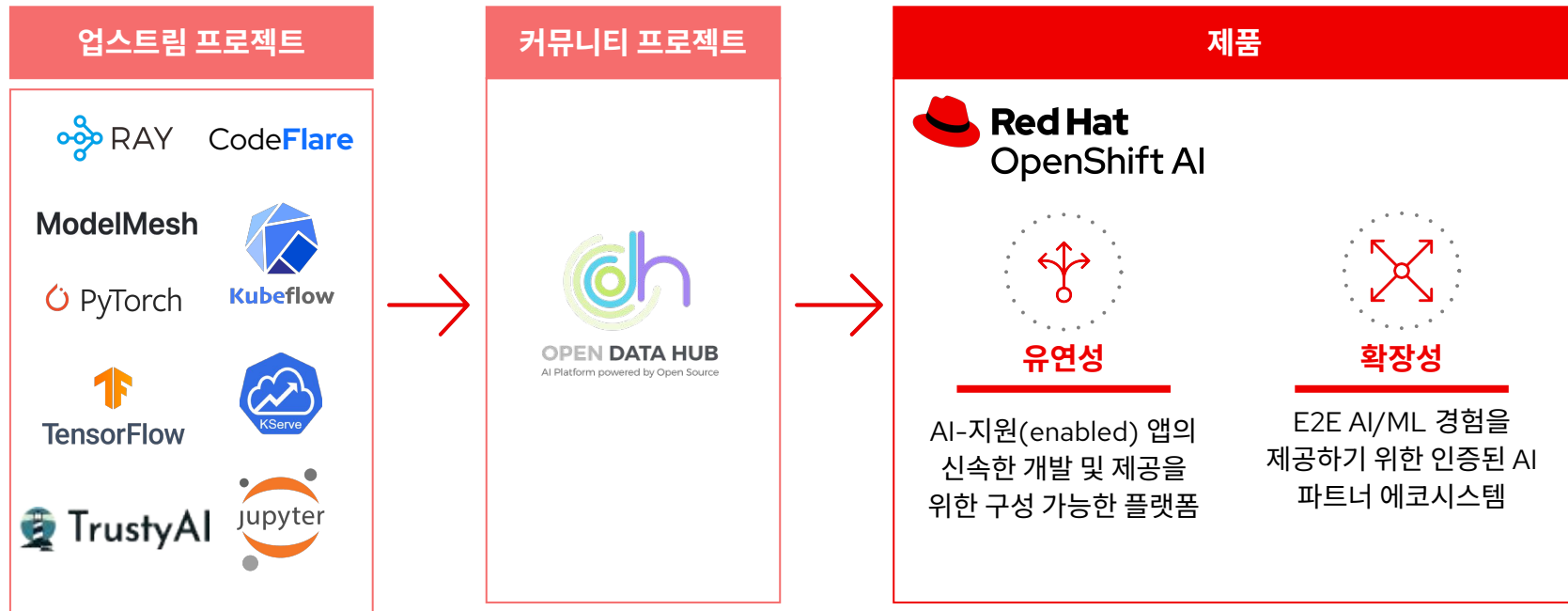
기업 전체에 걸쳐 모델의 튜닝, Alignment (Manual 혹은 InstructLab), 검증 및 전달을 위한 모델 파이프라인을 만듭니다.



리소스 관리 및 최적화

튜닝 및 추론에 사용할 리소스를 최적화합니다. 사용자 간 리소스, 프로젝트 및 모델을 공유합니다.

Red Hat의 AI/ML 엔지니어링은 100% 오픈소스입니다.



Red Hat OpenShift AI 기능 업데이트

하이브리드 클라우드 환경에서 MLOps 플랫폼을 제공하는 OpenShift AI에 대한 몇 가지 신규 기능을 발표되었습니다. 주로 인프라 모델 지원과 AI 실행 환경 확장을 통해, 고객의 요구에 맞는 환경에서 다양한 AI 모델을 개발하고 실행하는 것이 가능하게 되었습니다.



MLOps on Open Hybrid Cloud

분산학습

OSS 분산 학습 프레임워크인 Ray와 CodeFlare를 통해 멀티 노드를 사용한 학습 가능
- Job 관리에는 Kueue를 사용



LLM 서빙

LLM을 위한 서빙 런타임으로 vLLM을 지원
KServe와 결합하여 유연한 확장이 가능



HW 가속화

멀티 벤더의 가속 디바이스를 사용하고 모델 학습 및 추론을 위해 컨테이너에 연결



엣지

클라우드, 온프레미스 이외의 모델 실행 환경으로 공장(factory) 등 Near Edge 환경을 선택 가능

← 파운데이션 모델 향상(Enhancements) →

← AI 실행 환경을 위한 옵션 확장 →

Red Hat OpenShift AI 소프트웨어 스택



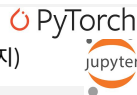
MLOps on Open Hybrid Cloud

학습
Model Training

추론
Model Inference

실험용 학습 환경

- JupyterLab
- PyTorch/TensorFlow/Scikit-learn (컨테이너 이미지)



학습 작업 관리

- Kueue



분산 학습 환경

- Ray
- CodeFlare



RAY



파이프라인

- KubeFlow Pipelines



리소스 관리*

- Nvidia GPU Operator
- Intel Habana AI Operator
- AMD GPU Operator

제공되는 인퍼런스 API

- KServe
- OVMS



LLM 서빙

- TGIS Serving
- vLLM



AI 모델 모니터링

- TrustyAI
- Prometheus



파트너와의 협업

애플리케이션 영역부터 하드웨어 영역까지 OpenShift AI와 파트너사 간의 광범위한 협력이 발표되었습니다. 향후 각 기업의 솔루션들을 OpenShift AI와 결합해 이전보다 훨씬 더 많은 고객의 요구를 충족할 예정입니다.

LLM

Stability AI

OpenShift AI에서 Stability AI가 제공하는 모델 서빙

RAG

Elastic

Elasticsearch를 VectorDB로 사용하여 RAG 구현

GPU Scheduling

Run:ai

Run:ai를 통한 GPU 스케줄링 및 활용 최적화

Inference

NVIDIA

NVIDIA NIM의 마이크로 서비스 기반 AI인퍼런스 환경 배포

Hardware Acceleration

Intel

데이터 센터에서 옛지까지 다양한 인텔 디바이스 사용 가능

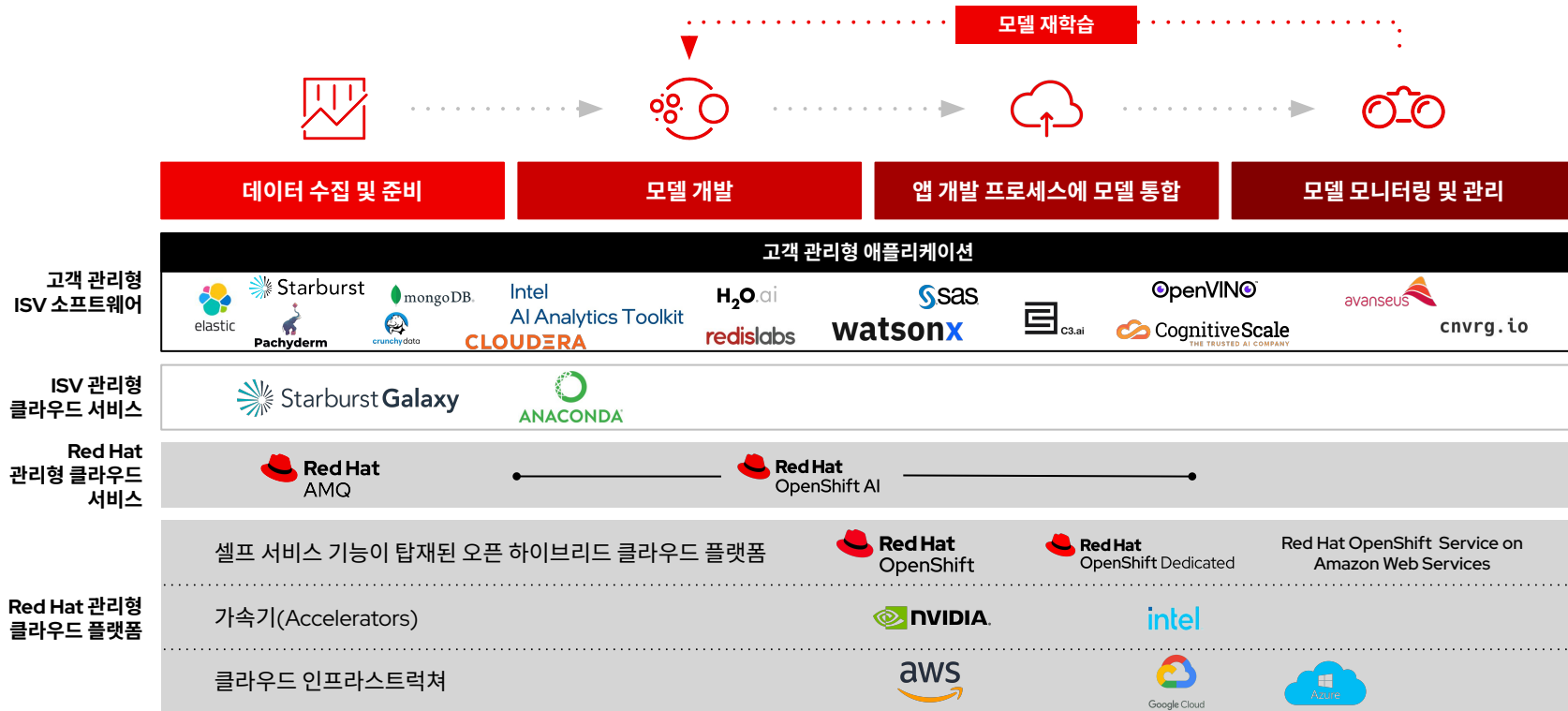
Hardware Acceleration

AMD

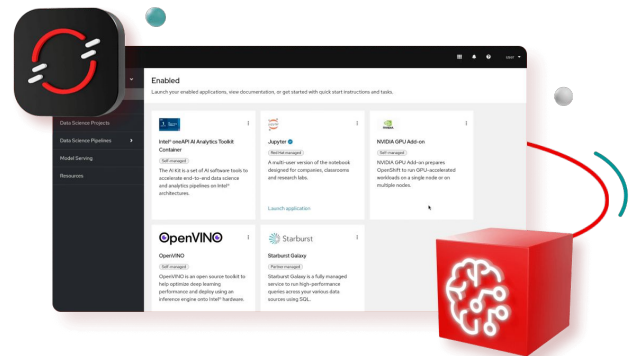
AI 훈련 및 인퍼런스를 위해 OpenShift 클러스터에서 AMD GPU 활용



OpenShift AI 구성요소

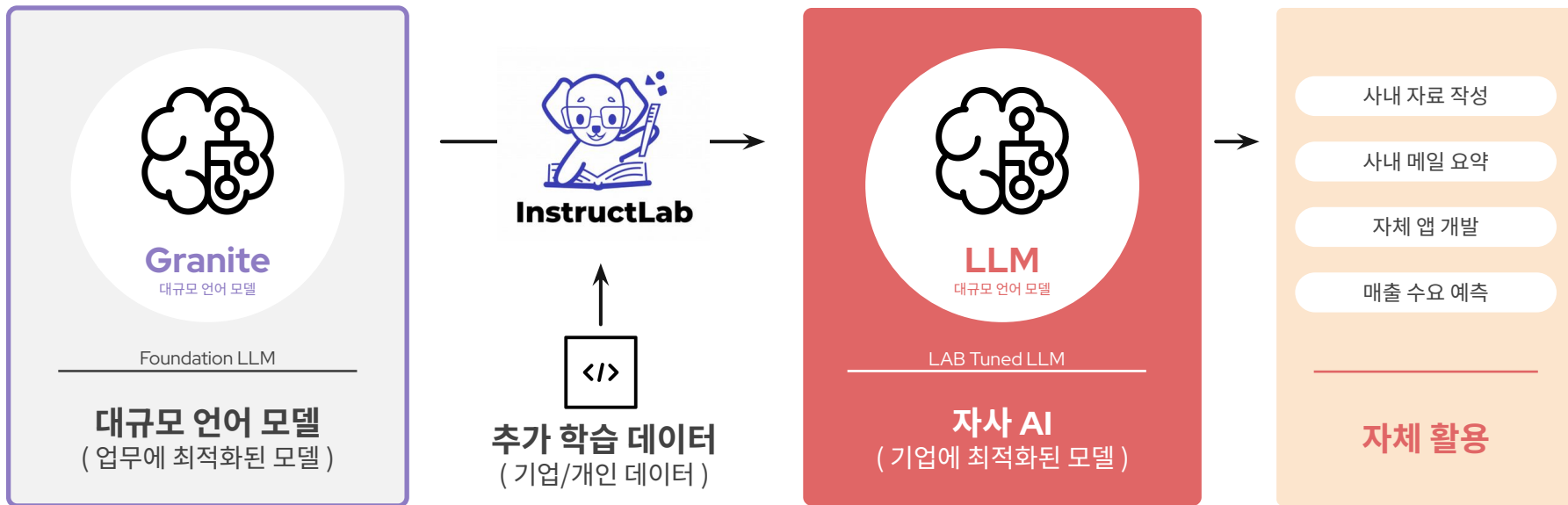


Wrap-up



오픈소스를 통한 AI의 대중화

누구나 모델을 최적화할 수 있는 플랫폼을 제공합니다



End-to-End AI 개발 루프 제공

로컬 환경에서 AI를 작게 시작하고 엔터프라이즈 규모로 코드를 확장하세요



STEP 1

로컬 Model을 실행하고 Code를 작성.

소규모 데이터 세트에서 제한된 데스크톱 규모 트레이닝 방법(qlora)으로 학습하고 실험.

 Laptop / desktop



Red Hat
Enterprise Linux AI

STEP 2

Full Synthetic Data 생성, Teacher 및 Critic Model을 사용한 상용 등급의 Model training. 스크립트 가능한 기본 요소에 초점을 맞춘 툴링.

 Server / VM



Red Hat
OpenShift AI

STEP 3


Kubernetes 확장, 자동화, MLOps 서비스 등 모든 기능을 활용하여 RHEL AI와 마찬가지로 상용 수준의 Model을 학습시킴.

 Cluster

Thank you

Red Hat is the world's leading provider of enterprise open source software solutions. Award-winning support, training, and consulting services make Red Hat a trusted adviser to the Fortune 500.

 [linkedin.com/company/red-hat](https://www.linkedin.com/company/red-hat)

 [youtube.com/user/RedHatVideos](https://www.youtube.com/user/RedHatVideos)

 [facebook.com/redhatinc](https://www.facebook.com/redhatinc)

 twitter.com/RedHat