MakinaRocks

ATTENTION
AI Engineered for the Real World

Beyond MLOps

Hyperconnectivity의 미래를 열다

심상우 CTO, MakinaRocks

AI 도입을 통한
생산성 향상

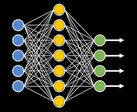
AI 운영 체계의 발전

MLOPS

LLMOPS

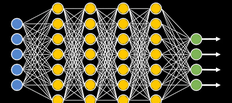
Compound AI System

Simple Neural Network



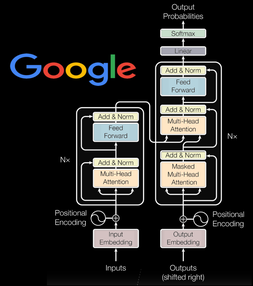
Machine Learning

Deep Learning Neural Network



Deep Learning

Google



Large Model (Transformer)

OpenAI

Foundation Model

AI Agent

2024

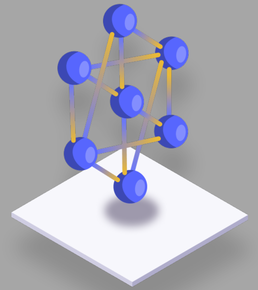
2012 2017 2022

AI 기술 및 운영 체계의 발전

LLM의 등장
AI 모델 학습과 성능 개선
AI의 실질적인 응용

단일 AI 모델 운영을 위한

단일 시스템



VS

복수 AI,
여러 구성요소가 포함된

복합 시스템

Compound AI System

A system that tackles AI tasks using multiple interacting components, including multiple calls to models, retrievers, or external tools.

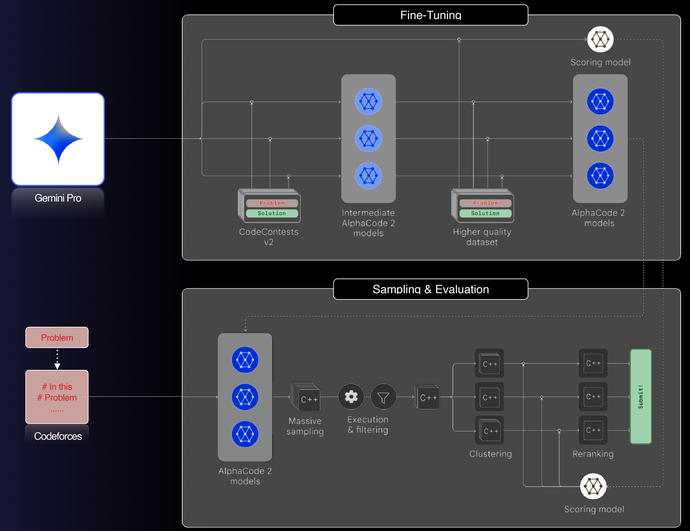
... most of the state-of-the-art implementations are systems with multiple specialized components.

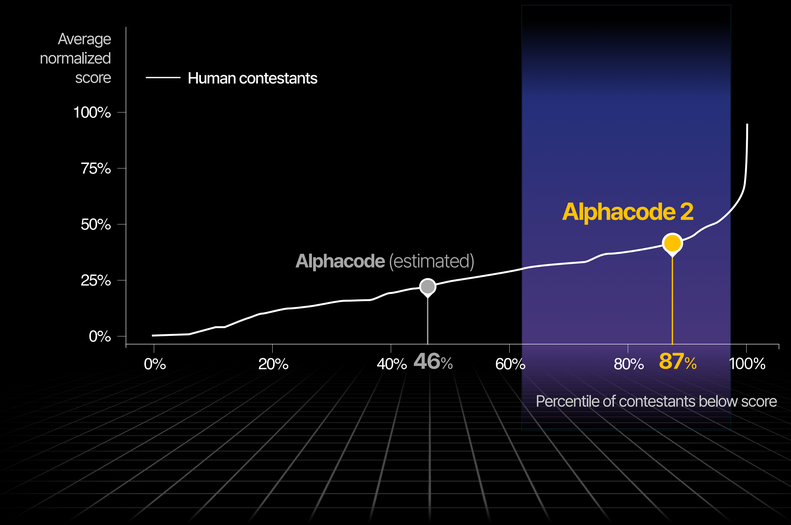
BAIR(Berkeley AI Research)

AI 모델, 검색기, 외부 도구를 여러차례 호출해
여러 구성요소의 상호작용을 통해
AI 작업을 처리하는 시스템

... SOTA 구현체는 대부분 여러 개의 특화된 컴포넌트가 결합된 시스템으로 구성

AlphaCode2 의 복합 AI 시스템 구성도





Result

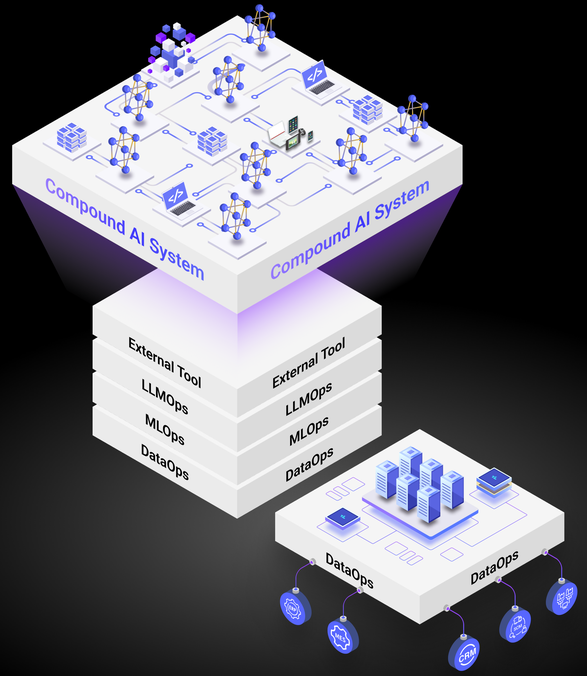
- LLM 포함 복합 AI 모델과 샘플링/필터링 메커니즘을 결합
- 기존 AlphaCode 대비 약 2배의 성능 향상
- 인간코더 상위 15% 이상의 점수 달성

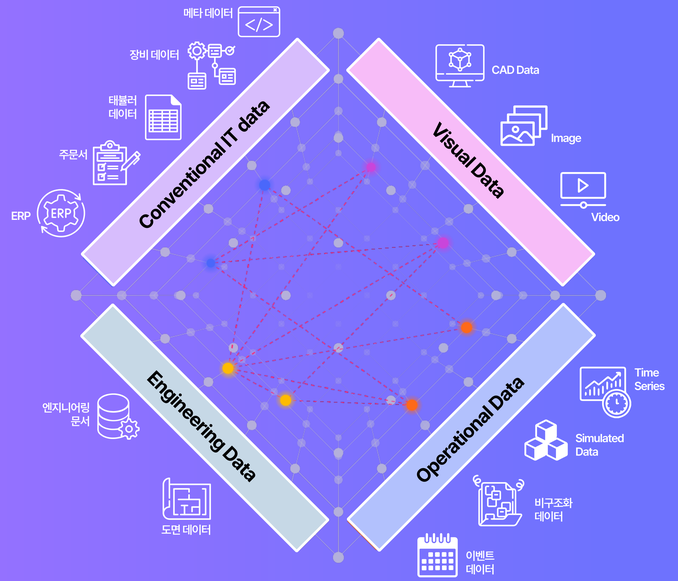
Compound AI System →

구성요소 #1

DataOps

기업 고유의 시스템에서 발생하는 다양한 멀티모달 데이터를
연결, 수집, 변환, 전처리(어노테이션 등) 하여 AI 적용을 준비



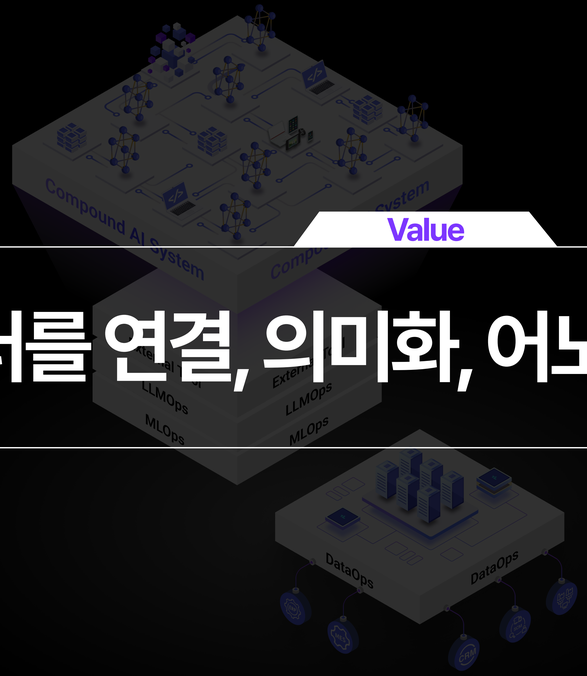


Compound AI System →

구성요소 #1

DataOps

기업 고유의 시스템에서 발생하는 다양한 멀티모달 데이터를
연결, 수집, 변환, 전처리(어노테이션 등) 하여 AI 적용을 준비



Value

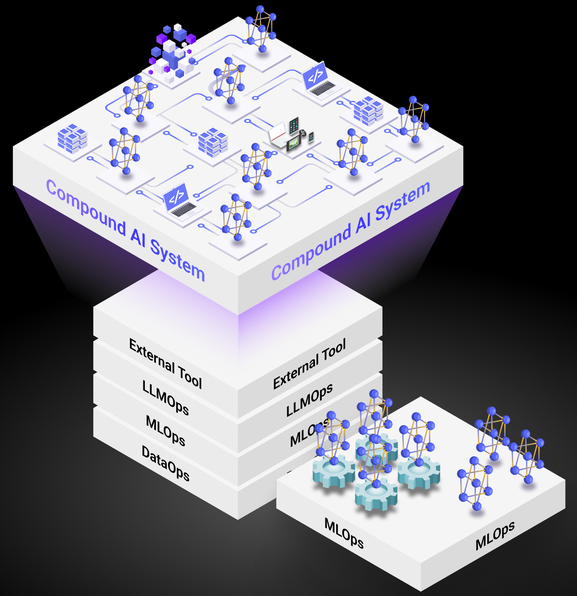
다양한 유형의 데이터를 연결, 의미화, 어노테이션하여 AI 최적화에 기여

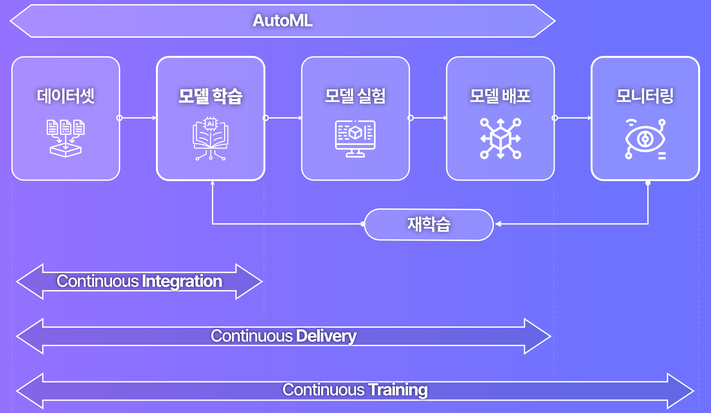
Compound AI System →

구성요소 #2

MLOps

AutoML과 MLOps를 통해 AI 모델의 개발부터 배포에 이르는
전 과정을 자동화하여 반복 작업을 최소화하고,
자율적으로 운영되는 AI 시스템을 구축





Compound AI System →

구성요소 #2

MLOps

오토ML과 MLOps를 통해 AI 모델의 개발부터 배포에 이르는
전 과정을 자동화하여 반복 작업을 최소화하고,
자율적으로 운영되는 AI 시스템을 구축

Value

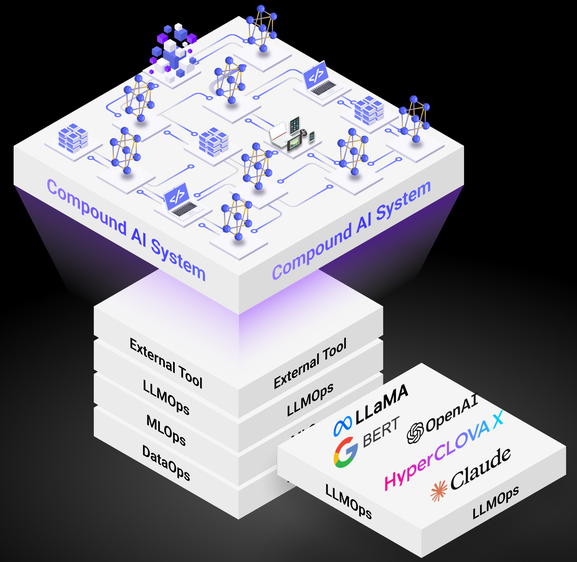
다양한 AI 모델의 개발-배포, 안정적인 유지보수를 보장

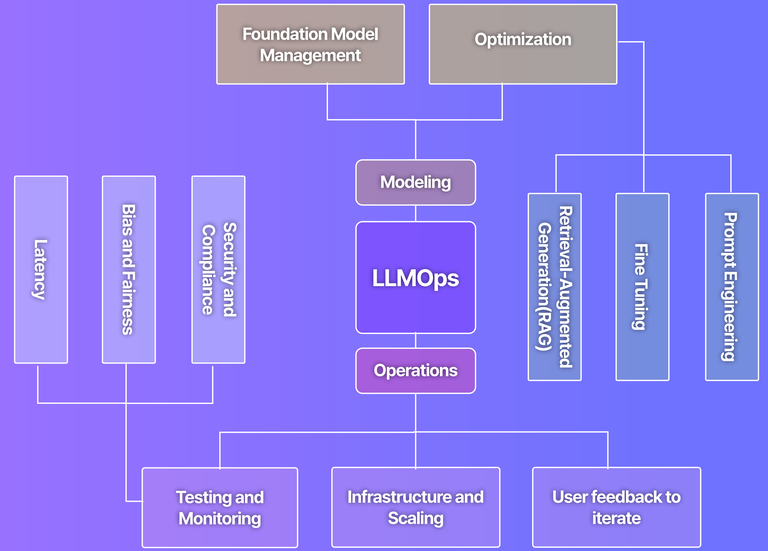
Compound AI System →

구성요소 #3

LLMOps

다양한 LLM 모델을 쉽고 안정적으로 배포하고 운영하며,
특화된 런타임과 서빙 인프라의 자원 최적화를 지원



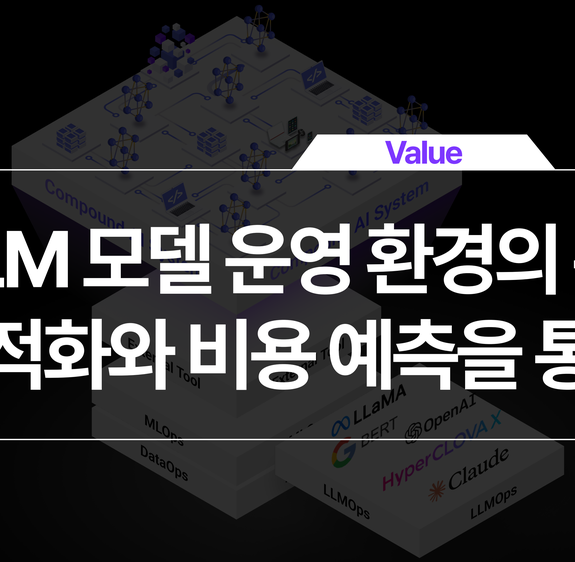


Compound AI System →

구성요소 #3

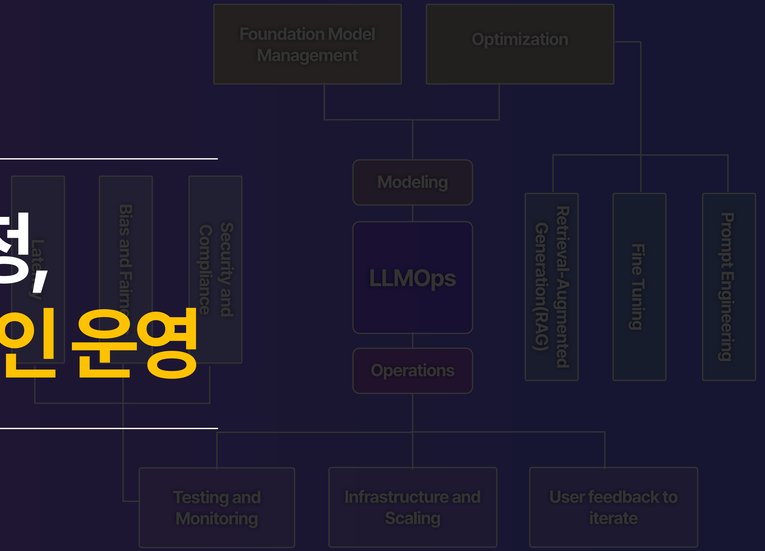
LLMOps

다양한 LLM 모델을 쉽고 안정적으로 배포하고 운영하며,
특화된 런타임과 서빙 인프라의 자원 최적화를 지원



Value

LLM 모델 운영 환경의 손쉬운 설정, 성능 최적화와 비용 예측을 통한 효율적인 운영

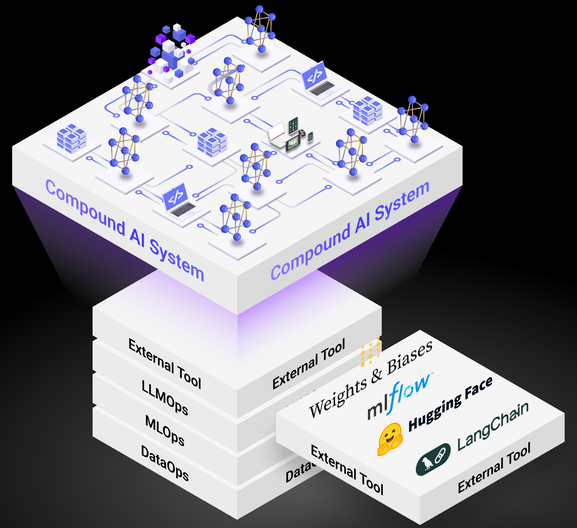


Compound AI System →

구성요소 #4

External Tool 연동

AI 및 대형 언어 모델(LLM)의 개발과 운영을 위해
다양한 외부 툴과의 매끄러운 연동을 지원하여,
보다 유연하고 확장 가능한 AI 환경을 제공



databricks
jupyter
Weights & Biases
hadoop
PC
mlflow
Power BI
snowflake
TensorFlow
Grafana
Amazon ECR
MINIO
PyTorch
HARBOR
aws
FriendliAI
Google Cloud
Hugging Face
docker

Compound AI System →

구성요소 #4

External Tool 연동

AI 및 대형 언어 모델(LLM)의 개발과 운영을 위해
다양한 외부 툴과의 매끄러운 연동을 지원하여,
보다 유연하고 확장 가능한 AI 환경을 제공

Value

시스템의 기능을 확장하여 실질적인 문제 해결 능력 강화

databricks
jupyter
Weights & Biases
hadoop
PC
mlflow
Power BI
TensorFlow
Grafana
Amazon ECR
MINIO
PyTorch
HARBOR
aws
FriendliAI
Google Cloud
Hugging Face
docker

Hyperconnected AI Platform,
Runway

기업의 문제를 해결하기 위해
AI모델, 멀티모달 데이터, 그리고 기업 시스템을
유기적으로 연결하는 AI 플랫폼

Connect Your AI

연결성

멀티모달 데이터와 AI 모델, 그리고 기업 고유 시스템을 연결합니다.
사용자가 실시간으로 AI 인사이트를 얻을 수 있는
통합 운영 환경을 제공합니다.

01. 데이터

학습 및 모델 활용을 위한
여러 데이터 소스들의 연결

DB, Object Storage
텍스트, 이미지, 비디오, 센서

02. 멀티 모델

데이터의 형태에 따라
맞춤형으로 적용되어
복합적인 문제 해결

LLM, 텍스트-이미지 모델,
최적화, 예측 ...

03. 인프라

개발 및 서빙을 위해 필요한
다양한 인프라의 연결

온프레미스, 클라우드, 하이브리드,
폐쇄망 환경

04. 모든 장비

각종 장비에서
서빙할 수 있도록
On-device AI 지원

EdgeOps

Automate Your AI

자동화

복잡했던 AI 프로세스가 단순해집니다.
클릭만으로 맞춤형 AI 모델이 자동 생성되고 원하는 방식으로
간편하게 배포하고 지속 운영할 수 있습니다.

01. Annotation

최적의 데이터셋
구성을 위한 AI를 접목한
데이터 전처리

02. AutoML

최적의 피처를
자동으로 선택하고,
가장 적합한 AI 모델을 생성

03. CI/CD

모델을 지속적으로
통합하여 자동으로
테스트하고 배포

04. CT

새로운 데이터를
기반으로 모델을
지속 재학습 및 성능 유지

Govern Your AI

투명하고 안정된 운영

최고 수준의 보안과 자원 모니터링으로 데이터를 보호하고,
모든 시스템 활동을 투명하게 추적합니다.
복잡한 설정 없이도 AI의 안정성과 규정 준수를 확보할 수 있습니다.

01.
엔터프라이즈급
보안

- 역할 기반 접근 제어 및 관리
- 액세스 토큰 기반 제어 및 관리

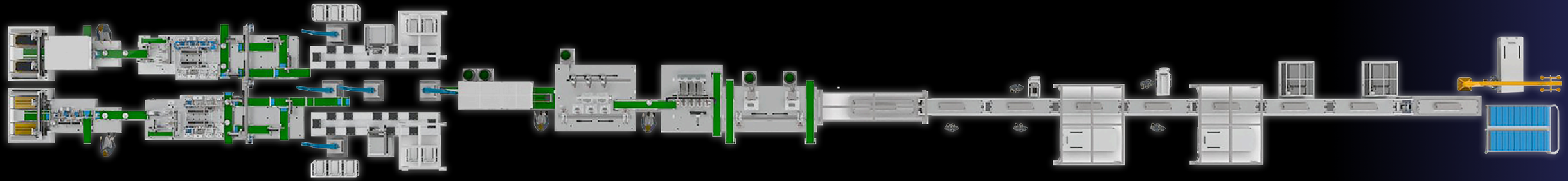
02.
실시간
모니터링

- 사용자, 프로젝트 워크스페이스 별 자원 사용 모니터링
- AI 모델 성능 모니터링

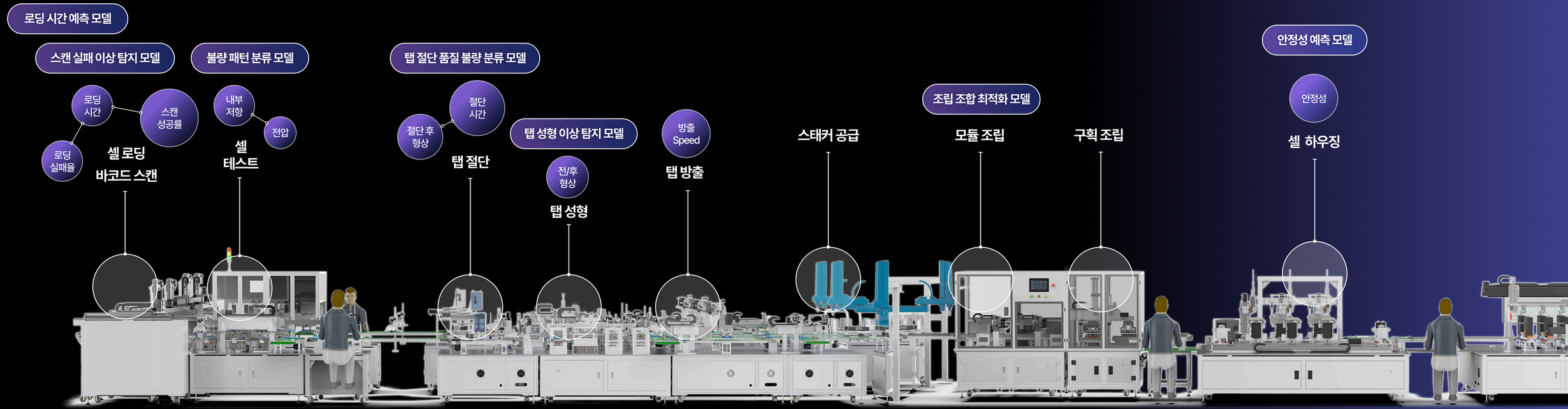
03.
재현성 및 추적성

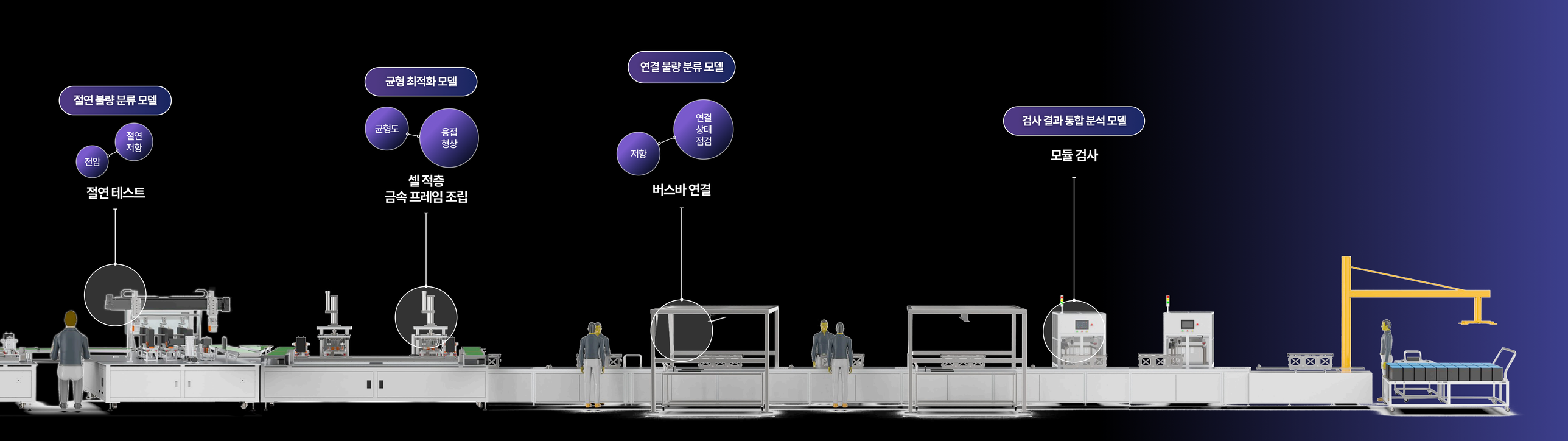
- 데이터 버저닝
- 데이터, 실험, 모델, 추론 서비스 리니지 관리
- 프로젝트 백업 및 복원

Compound AI System in Action



소프트팩 배터리의 자동생산 라인


로딩 시간 예측 모델
스캔 실패 이상 탐지 모델
로딩 시간
스캔 성공률
로딩 실패율
셀 로딩 바코드 스캔
불량 패턴 분류 모델
내부 저항
전압
셀 테스트
탭 절단 품질 불량 분류 모델
절단 시간
절단 후 형상
탭 절단
탭 성형 이상 탐지 모델
전/후 형상
탭 성형
방출 Speed
탭 방출
스태커 공급
조립 조합 최적화 모델
모듈 조립
구획 조립
안정성 예측 모델
안정성
셀 하우징


절연 불량 분류 모델
절연 저항
전압
절연 테스트
균형 최적화 모델
균형도
용접 형상
셀 적층
금속 프레임 조립
연결 불량 분류 모델
연결 상태 점검
저항
버스바 연결
검사 결과 통합 분석 모델
모듈 검사

AI 플랫폼 퍼스트 무버, Runway

GYEONGBUK HYBRID TECHNOLOGY INSTITUTE

KIDI Korea Insurance Development Institute

K water

Runway 덕분에 AI 지식이 없어도 복잡한 AI 모델을
손쉽게 재학습하고 배포할 수 있었습니다.
다양한 환경에 신속히 맞춤 통합이 가능해 운영 효율성이 크게 개선되었습니다.

HYUNDAI

SAMSUNG ELECTRO-MECHANICS

Runway는 AI 워크플로우의 핵심입니다.
데이터 라벨링부터 배포까지 모든 과정이 한 플랫폼에서 연결되어
작업 시간이 80% 단축되었고, 생산성이 크게 향상되었습니다.

Agency for Defense Development

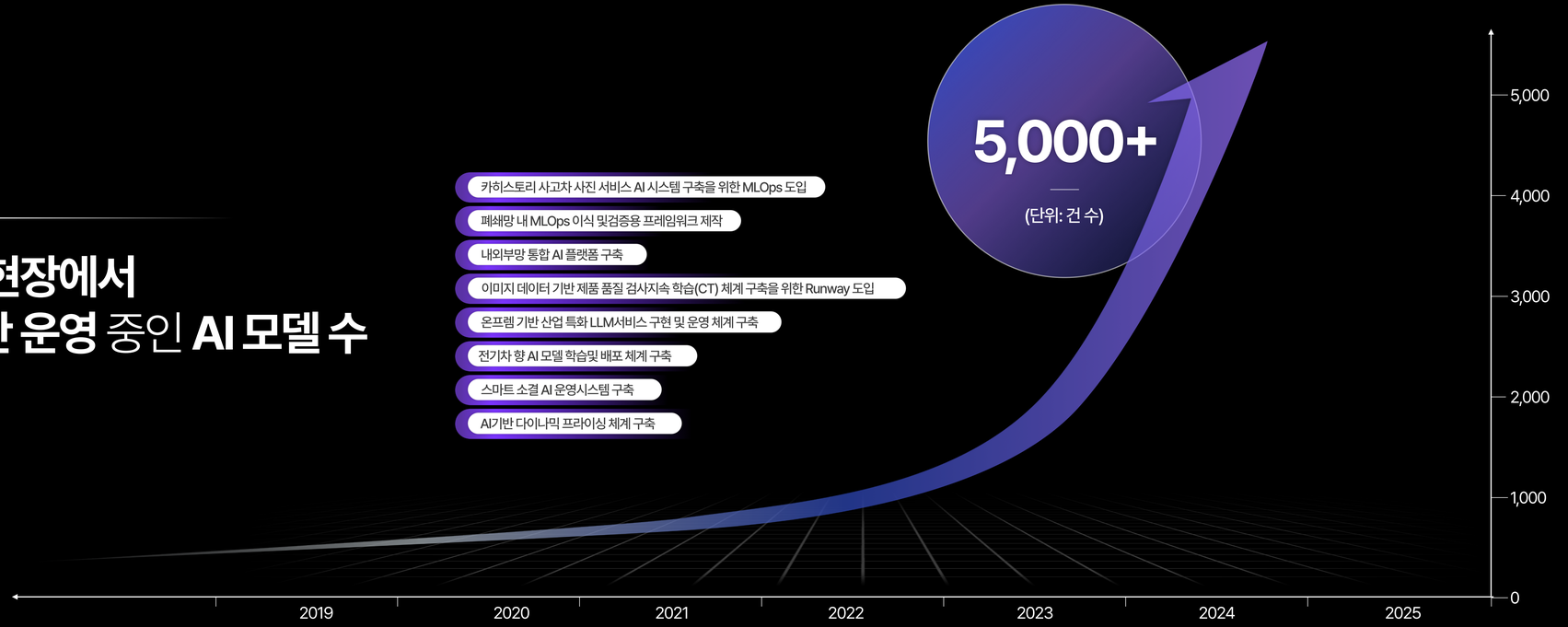
이연 FnC

Runway의 데이터 파이프라인과 모델 최적화 덕분에
프로세스 효율이 크게 개선되었습니다. Private Cloud에서도
안정적 운영이 가능해 IT 부담을 줄이면서 AI 성능을 최적화할 수 있었습니다.

차원이 다른 1등 플랫폼,
Runway

산업 현장에서 실시간 운영 중인 AI 모델 수

- 카히스토리 사고차 사진 서비스 AI 시스템 구축을 위한 MLOps 도입
- 폐쇄망 내 MLOps 이식 및검증용 프레임워크 제작
- 내외부망 통합 AI 플랫폼 구축
- 이미지 데이터 기반 제품 품질 검사지속 학습(CT) 체계 구축을 위한 Runway 도입
- 온프렘 기반 산업 특화 LLM서비스 구현 및 운영 체계 구축
- 전기차 향 AI 모델 학습및 배포 체계 구축
- 스마트 소결 AI 운영시스템 구축
- AI기반 다이나믹 프라이싱 체계 구축



엔터프라이즈를 위한 Runway,
이제 가볍게 시작하세요

Runway Lite

복잡한 인프라와 서버 설정 없이 간편한 AI 개발 시스템 도입

데이터셋과 실험 이력 관리 기능으로 동료와 효율적으로 협업

아이디어를 빠르게 검증해 AI 개발 생산성 향상

출시 알림 예약

THANK **YOU!**

MakinaRocks

Email Address
contact@makinarocks.ai

Website
www.makinarocks.ai