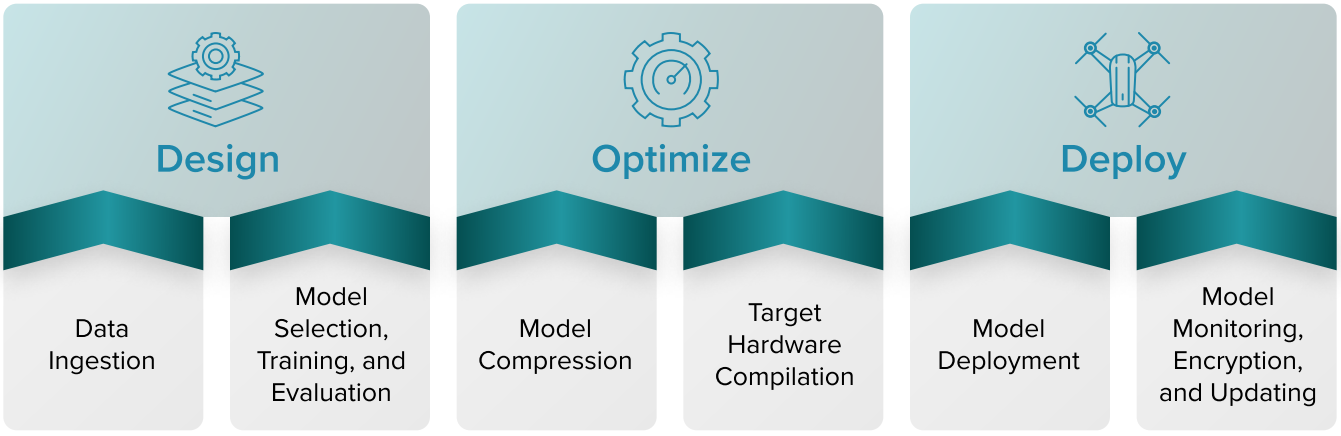
LatentAI

Mission-Critical Edge AI Solutions

Latent AI Efficient Inference Platform (LEIP)

Empowering warfighters in today's evolving battlefield demands interoperable and adaptable AI solutions capable of meeting diverse mission requirements. Providing trusted edge AI products and services, Latent AI empowers the DoD to seamlessly integrate and scale AI into all aspects of missions and operations. Our solutions are built with modular architecture and streamlined workflows to allow users of all skill levels to customize their AI solutions for the edge and rapidly adapt deployments. LEIP simplifies the complex process of deploying AI at the edge with an all-in-one platform offering unprecedented efficiency, cost savings, and scalability.



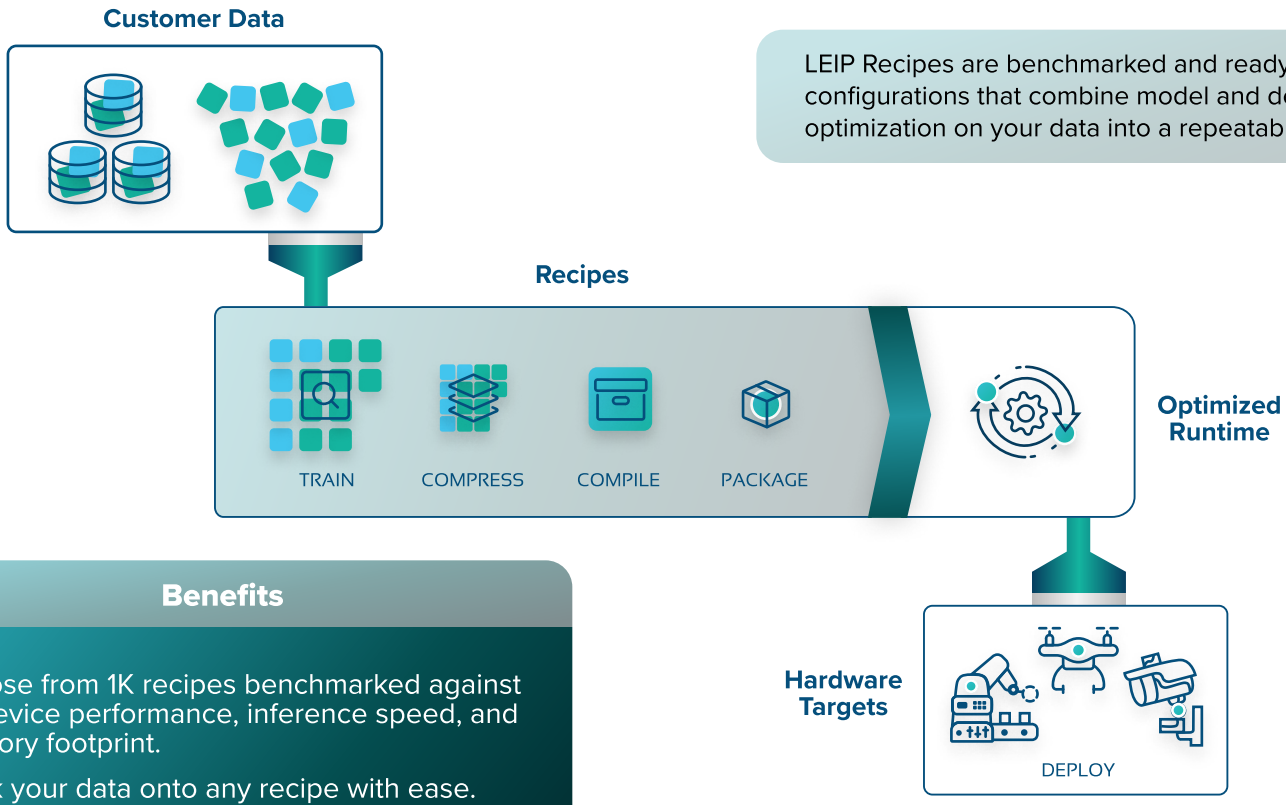
LEIP streamlines AI development with modular tooling and automation so that users of all skill levels can rapidly customize and adapt their AI solutions to the edge.

Benefits

- Streamline the entire ML pipeline for speed, consistency, and scale.
- Integrate with the tools and applications you need.
- Rapidly adapt AI to changing mission parameters with a scalable, repeatable process.
- Port AI across hardware platforms and environments.
- Ensure AI model security is built in.

Contact **info@latentai.com** or scan QR code to visit our website

Design | Predictable, effortless ML every time

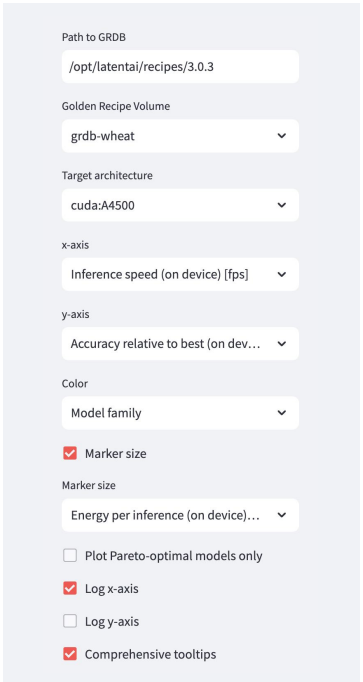


LEIP Recipes are benchmarked and ready-to-execute configurations that combine model and device optimization on your data into a repeatable process.

Benefits

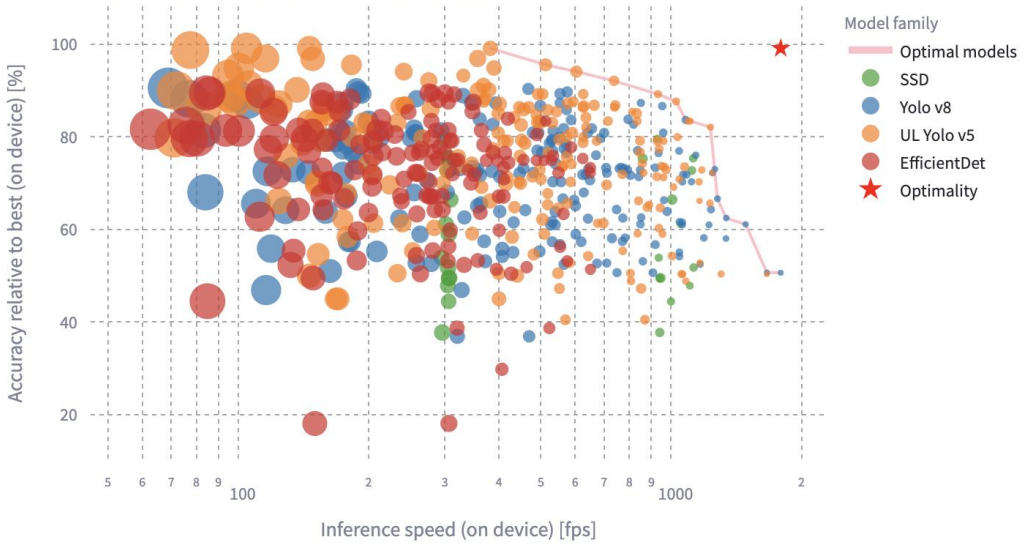
- Choose from 1K recipes benchmarked against on-device performance, inference speed, and memory footprint.
- Hook your data onto any recipe with ease.
- Analyze size, accuracy, and power trade-offs interactively and meet your exact criteria.
- Target multiple hardware platforms without restarting the design process.
- Reuse a modular pipeline to apply different models and/or hardware to the same dataset.

Most teams need help getting from idea to application. LEIP Design has a powerful visualizer that takes the guesswork out of choosing the right model-hardware combination for your project. Leveraging our deep expertise in optimization and performance benchmarking, you can effortlessly select the perfect fit for your needs.



LEIP Design

552 Recipes



Contact **info@latentai.com** or scan QR code to visit our website

Optimize | AI optimization simplified

Intuitive for Beginners, Powerful for Experts

Deploying AI to the edge is hard. Optimizing models for the hardware of your choice is time consuming and requires in-depth hardware knowledge. LEIP Optimize is a set of tools that automates rapid hardware and software optimization of ML models without requiring hardware expertise.

Optimize supports models from popular frameworks (Tensorflow, PyTorch, ONNX, etc.). It supports computer vision models and most models with dynamic tensors such as transformers, as well as the majority of models on Hugging Face.

Bring your own model or data

Model Optimization

Speed up model predictions via quantization and shrink to run in lower bit precision (INT8) while preserving accuracy.

Hardware Optimization

Convert an ML model to a single portable file which can leverage hardware acceleration libraries via a standardized API.

Features

- **Agnostic Model Ingestion**
  Bring your own model to a framework capable of ingesting models from various sources into a unified representation.
- **Forge**
  Simplifies and streamlines the processes of compiling, optimizing, and quantizing machine learning models, making these complex tasks accessible to a wider range of users.
- **Direct graph manipulation**
  Enables experts to debug and modify incompatible models for specific hardware.

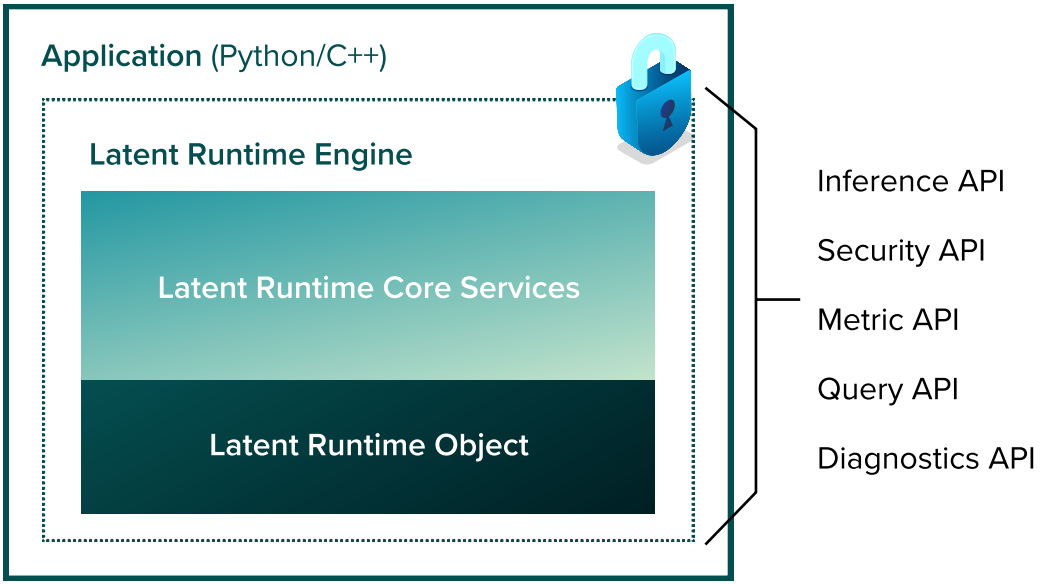
Benefits

- Automate model-hardware optimization for rapid prototyping and accelerate deployment.
- Target hardware without in-depth hardware expertise.
- Integrate into your current ML environment to maintain developer familiarity and tooling flexibility.
- Rapid prototype (test and evaluate) optimization for various model-hardware combinations.
- Optimize and encrypt multiple models for one or more hardware targets.
- Script optimization and compilation jobs for reusability and automation.
- Available on premises (SDK).

Contact **info@latentai.com** or scan QR code to visit our website

Deploy | Secure, easy-to-maintain runtime for your edge device

Once quantized and compiled, the Latent Runtime Engine (LRE) offers a standardized runtime engine for edge AI that is frictionless to deploy and maintain.



Mean time to update
Demonstrated as much as 18x reduction in model update time.

On device performance
Demonstrated 20% power savings after optimization on target hardware.

Inference speed
66% faster inference speed on an edge device (Jetson NX) than Sagemaker.

Features

- **Standardized data format** for your model to be optimized and ported to multiple hardware platforms. You can move from platform to platform with agnostic hardware support.
- **Update or replace** seamlessly with no changes to your application.
- **Measure performance** during deployment with real-time diagnostic metrics.
- **Track and secure** with a UUID so you can deter unauthorized use or distribution of models. LRE can be encrypted to protect from theft or compromise.
- **Integrate into your application** easily with C++ and Python compatibility and APIs for easy third-party and application integration.

PROJECT OVERMATCH

U.S. ARMY
PROJECT LINCHPIN

U.S. AIR FORCE

TECH

DEFENSE INNOVATION UNIT

CDAO

AWARDABLE

About Latent AI

Latent AI, Inc. is a leading expert in edge AI, specializing in simplifying the complex process of implementing AI on any device. Established in 2018, Latent AI's cutting-edge developer platform is trusted by both government and commercial organizations looking to revolutionize their operations by harnessing the power of AI at the edge. Our tools empower developers to rapidly build secure, adaptive models, and seamlessly update them in the field or lab. For more information on how we help organizations create better and safer AI more quickly, please visit latentai.com.

Contact **info@latentai.com** or scan QR code to visit our website