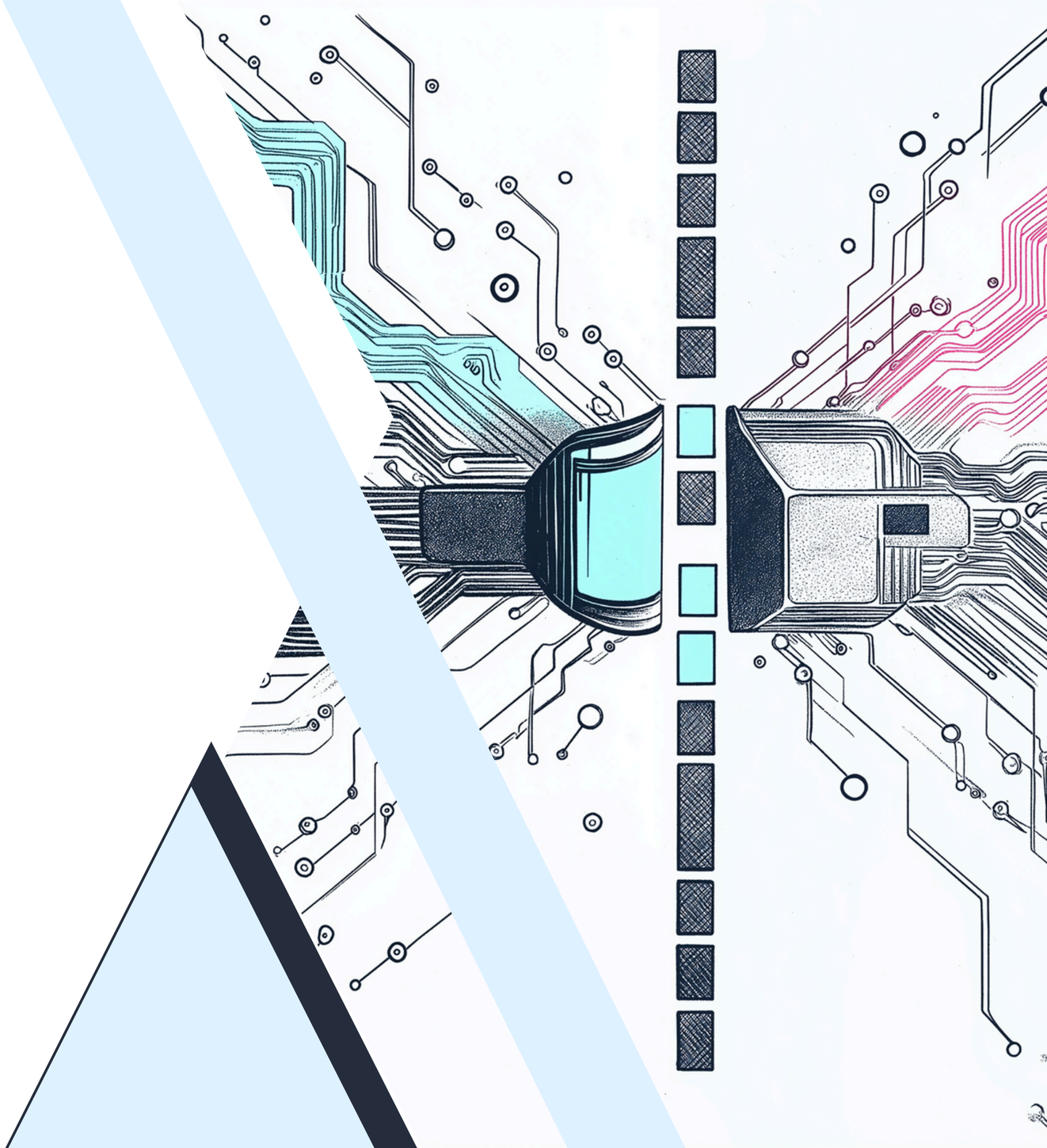
LatticaAI

HEAL

From Hardware to Private AI: Seamless Paths to Deployment



The Story of Why We Built HEAL

LatticaAI

Platform Vision

Privacy-Preserving Inference

Latency was too high

V1: PyTorch + GPU

Rapid prototyping with native acceleration

No native support for efficient mod-arithmetic:
redundant memory allocations, not int64 support, etc..

V2: Custom CUDA Kernels

Smarter memory management, kernel fusion

Low level operations like NTT became the bottleneck

V3: Custom FHE Chips

State-of-the-art FHE acceleration

Can't reimplement all high-level logic per chip

HEAL

Bridge to plug in custom hardware

What is HEAL

LatticaAI

HEAL is a production-focused integration suite

github.com/Lattica-ai/heal

What HEAL is

A hardware–software interface
Derived from our early PyTorch/GPU stack

A runtime for executing encrypted queries
Part of Lattica's FHE platform , built to execute queries across hardware backends

A dev and testing toolkit
Includes unit tests, example instruction transcripts, standalone CPU backend, and a standalone runtime tool

What HEAL is not

Not a new FHE library
HEAL wraps Lattica's FHE core, but only exposes low-level execution traces

Doesn't handle Enc, Dec, KeyGen
Those happen externally via the Lattica Query Client

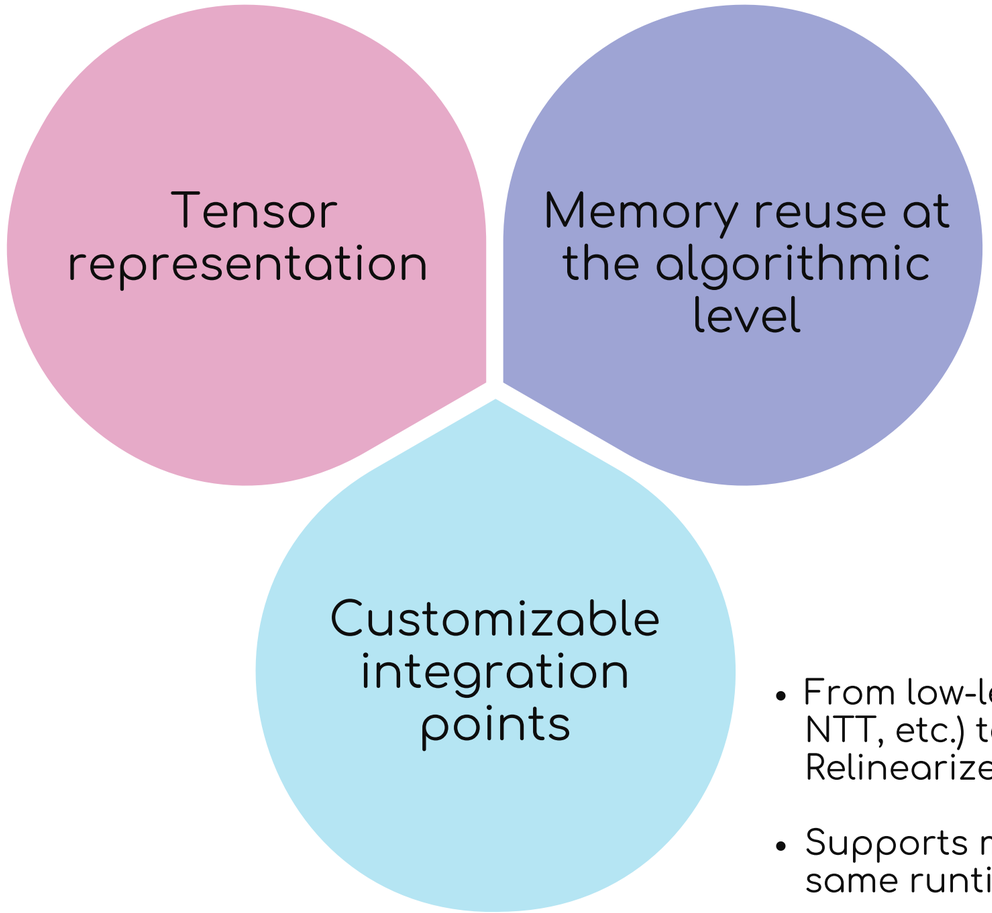
Not complicated
Minimal by design, built for real deployment and easy integration

Real Engineering Learnings

LatticaAI

Both memory I/O and compute can be bottlenecks, they must be addressed holistically

- Maximizes algorithmic parallelism potential
- Non-contiguous inputs and standard broadcasting



- Enables buffer reuse across ops
- works via pointers, strides and shape metadata
- Does not manage internal memory hierarchy

- From low-level primitives (modmul, decomp, NTT, etc.) to high-level op groups (KeySwitch, Relinearize, etc.)
- Supports multiple hardware variants in the same runtime graph

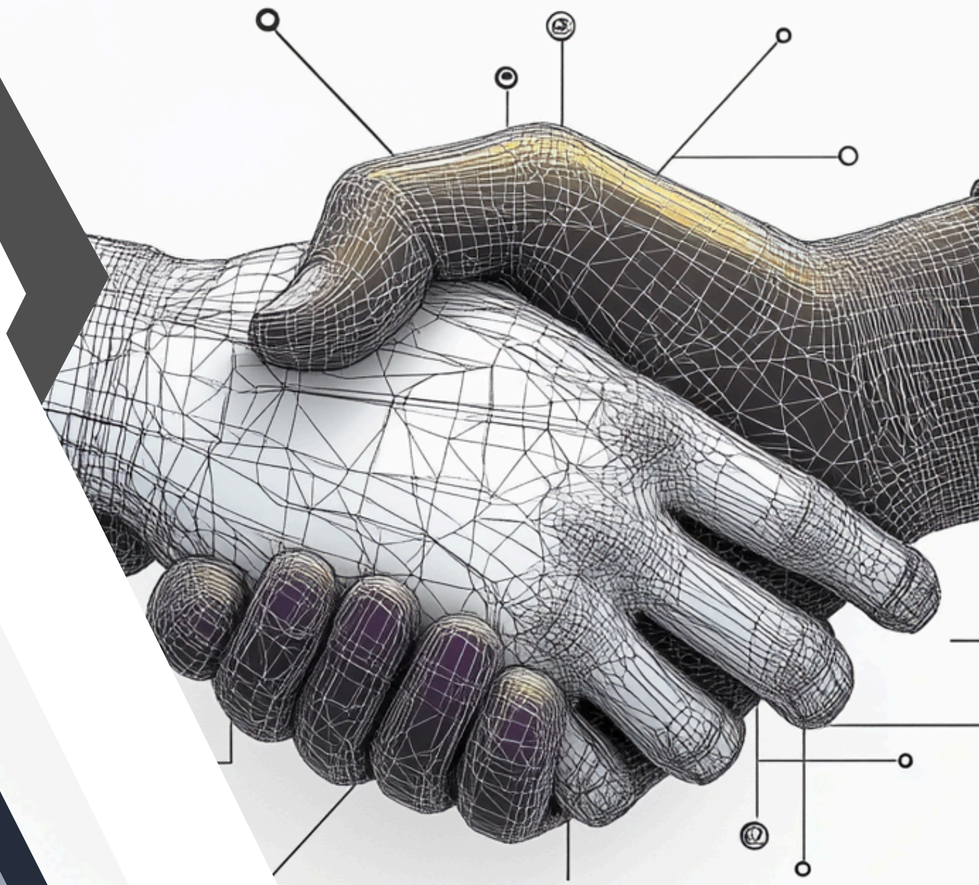
LatticaAI

Aligned Incentives

Hardware unlocks new use cases
Lower latency, higher throughput → enables deeper models and more verticals.

We offer go-to-market
AI workloads, Deployment infra, query APIs, usage analytics.

Plug-and-play, not partner-heavy
Single interface works across accelerators, cloud-agnostic.
We want to be easy to work with.



Thank You!

rotem@lattica.ai