

# Edge AI Deployment & Inference Optimisation

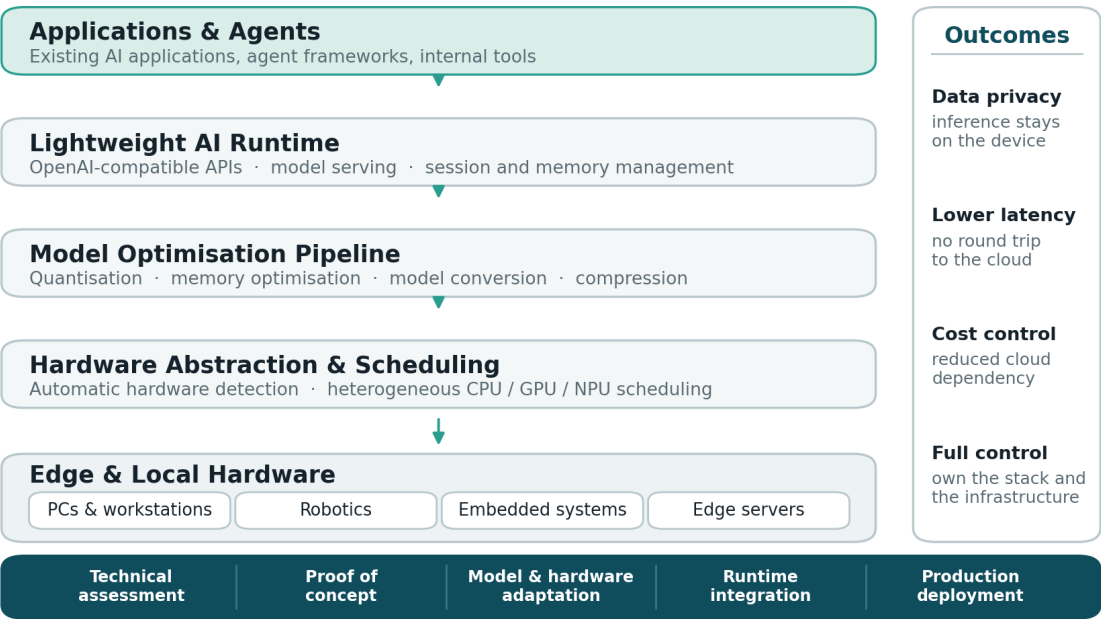
PolyhedronTech | Hong Kong | Belt and Road Summit 2026 - Deal-Making Platform

## Project summary

PolyhedronTech deploys and optimises large language models and multimodal AI on edge and local devices. The project delivers model optimisation, quantisation, heterogeneous CPU/GPU/NPU acceleration, runtime integration and private AI deployment for enterprises, device manufacturers and technology partners that need advanced AI to run efficiently outside the cloud.

## Edge AI Deployment & Inference Optimisation

PolyhedronTech | reference architecture for private, on-device AI



## Capability stack

- Model optimisation, quantisation and compression
- Model conversion and memory optimisation
- Automatic hardware detection
- Heterogeneous CPU / GPU / NPU scheduling
- Lightweight runtime with OpenAI-compatible APIs
- Integration with existing apps and agent frameworks

## Deployment scenarios

- Enterprise PCs and workstations
- Robotics and autonomous systems
- Embedded and constrained-memory devices
- On-premise and edge servers
- Regulated environments requiring data residency
- Sites with limited or no connectivity

### Engagement model

- 1 Technical assessment
- 2 Proof of concept
- 3 Model & hardware adaptation
- 4 Runtime integration
- 5 Production deployment

## What we are seeking at the summit

Deployment partners and system integrators; device manufacturers and hardware partners; channel partners for Mainland China, South-East Asia, the Middle East and Europe; and investors interested in edge AI infrastructure. Project location: Hong Kong and Mainland China. Project format: private project. Stage: investment and financing discussion.