

SPECIALIST SKILL / EDGE & LOCAL INFERENCE

Run the model where the decision happens.

Keep inference close to equipment, users or protected data when cloud latency, connectivity or data movement is the wrong fit. Select and optimize against the actual device envelope.

WHAT WE CAN BUILD FOR YOU

Inspection on an intermittently connected site

Optimize a model for a local device and design buffering and offline behavior around its sensing pipeline.

Value: The inspection workflow can continue when connectivity is interrupted.

Private inference on local hardware

Benchmark model quality, memory use and response time on candidate devices before packaging deployment.

Value: A hardware decision based on measured trade-offs, with less dependence on moving data off site.

WHAT THE ENGAGEMENT INCLUDES

- | Quantisation and compilation
- | Hardware selection
- | On-device pipelines
- | Offline operation

WHAT YOU TAKE AWAY

- | Optimised model artefacts
- | Benchmark report
- | Device deployment package

HOW WE MAKE THE VALUE VISIBLE

On-device latency / Power and memory use / Offline completion rate

GOOD REASONS TO START

Latency or power budgets exceeded / Data that cannot leave the site / Intermittent connectivity

ILLUSTRATIVE APPLICATIONS / SCOPED TO YOUR SYSTEMS

Let's scope the first useful step.

Bring the model, target hardware options and the latency, power or data-boundary constraint.

[Talk to BELTO](#)