SPECIALIST SKILL / AI INFRASTRUCTURE & AUTOMATION

Get useful work out of your AI compute.

Connect model serving, GPU scheduling and workflow integration so demand reaches the right capacity. Expose bottlenecks and unit cost instead of treating a growing infrastructure bill as inevitable.

WHAT WE CAN BUILD FOR YOU

A shared model-serving platform

Route inference jobs through queues, schedule GPU capacity and scale serving resources against workload demand.

Value: Teams can share expensive resources with clearer priorities and fewer unexplained delays.

From model result to business workflow

Connect serving outputs to downstream systems with controlled handoffs and observable execution.

Value: A model becomes part of a usable process rather than another manual step.

WHAT THE ENGAGEMENT INCLUDES

- Serving architecture
- GPU scheduling
- Queueing and autoscaling
- Workflow integration

WHAT YOU TAKE AWAY

- Serving platform
- Scheduling configuration
- Integration components

HOW WE MAKE THE VALUE VISIBLE

GPU utilization / Queue delay / Serving cost per request

ILLUSTRATIVE APPLICATIONS / SCOPED TO YOUR SYSTEMS

Let’s scope the first useful step.

Bring workload traces, concurrency needs and the business process waiting for the result.

Talk to BELTO