LMP Local Inference & IOL Agent Summary

LMP Consulting — Local Inference & IOL Agent Summary

Prepared: 2026-03-20
Author: Koda 🐾 (AI Agent)
Company: LMP Consulting s.r.o., Trenčín, Slovakia

---

1. GPU Inference Cluster

Hardware

Node	Hostname	CPU	RAM	GPU	Storage	Network
spark	10.1.2.150	GB10 Grace Blackwell (20C)	128GB LPDDR5x	NVIDIA Blackwell	1TB SSD PCIe 4.0	10GbE + 2× 200GbE QSFP
dark	10.1.2.155	GB10 Grace Blackwell (20C)	128GB LPDDR5x	NVIDIA Blackwell	4TB SSD PCIe 5.0	10GbE + 2× 200GbE QSFP

- **Interconnect:** 200GbE InfiniBand (RDMA), NCCL v2.28.9-1 compiled with Blackwell support
- **Network:** All hosts on 10.1.2.0/24

Active Inference Endpoints

Endpoint	Host	Model	Backend	Context	Status
vLLM (multi-node)	spark:8355 (head) + dark (worker)	Nemotron-3-Super-120B (NVFP4)	vLLM + Ray	262K tokens	✅ Operational
Ollama	dark:11434	Qwen2.5-VL 7B (vision)	Ollama	—	✅ Operational
Ollama	dark:11434	EuroLLM-22B	Ollama	—	✅ Operational
Ollama	dark:11434	bge-m3 (embeddings)	Ollama	—	✅ Operational
Ollama			Ollama	—	

Endpoint	Host	Model	Backend	Context	Status
	dark: 11434	nomic-embed-text (embeddings)			✅ Operational

Proxy Layer (on OpenClaw host)

Proxy	Address	Target	Purpose
vLLM Super proxy	127.0.0.1:18357	spark:8355	Injects `enable_thinking: false` to suppress CoT leak
TRT-LLM dark proxy	127.0.0.1:18355	dark:8355	Strips unsupported OpenAI params
TRT-LLM spark proxy	127.0.0.1:18356	spark:8356	Strips unsupported OpenAI params
TRT-LLM load balancer	127.0.0.1:18350	dark + spark	Round-robin between TRT-LLM endpoints

> **Note:** TRT-LLM proxies (18355/18356/18350) are stale — GPUs are now allocated to vLLM. Cleanup pending.

Agent Routing Architecture

Agent Type	Model	Provider	Cost
Main agents (Koda, Catalyst, Nexus, Atlas)	Claude Opus/ Sonnet	Anthropic API	Paid
All subagents	Nemotron-3-Super-120B	Local vLLM cluster	Free

- Config key: `subagents.model` → `vllm-super/nvidia/NVIDIA-Nemotron-3-Super-120B-A12B-NVFP4`
- All 4 agents tested and confirmed routing correctly (2026-03-19)
- Nemotron leaks chain-of-thought when used as main model — safe only for subagent use

---

2. Infrastructure Services

Service	URL	Host
Portainer	https://portainer.lmphq.net	10.1.2.130:9443
Grafana	https://grafana.lmphq.net	10.1.2.130:3000
GitLab	https://gitlab.lmphq.net	—
Matrix	https://matrix.lmphq.net	10.1.2.188
Mattermost	https://matter.lmphq.net	10.1.2.188:8065

Service	URL	Host
Portal	https://portal.lmphq.net	10.1.2.188 (static)
CF Access Launcher	https://portal.lmp-ai.net	Cloudflare-hosted

Cloudflare tunnel (UUID: 1a379b4c) on MikroTik CCR2004, token-managed. Domain split: `lmp-ai.net` (public/CF) + `lmphq.net` (private/LAN Caddy).

---

3. IOL Agent Project — Predictive Maintenance

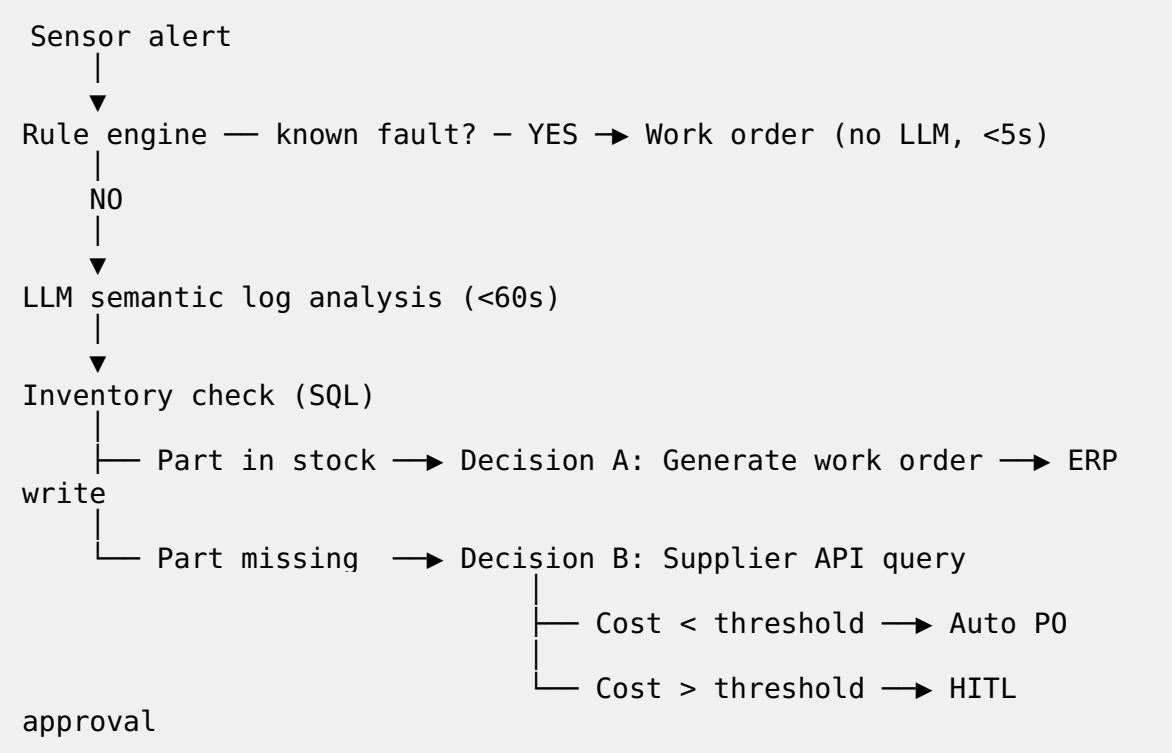
Overview

Projekt IOL Agent — Agentic AI system for predictive maintenance in manufacturing SMBs. Closed-loop automation: sensor anomaly → log analysis → inventory check → work order or purchase order → HITL approval above cost threshold.

Team

Agent	Role
Koda 🐾	Project Manager
Catalyst 🎯	Voice of the Customer
Nexus ⚙️	Technical Executioner
Atlas 🌟	Logging & Documentation

Architecture



Key design decisions: - Rule-based pre-filter handles 60-70% of known faults (no LLM, sub-5s) - LLM only for ambiguous/novel faults - On-premise only, no cloud dependency - €500 starting HITL threshold (adjustable) - Modular ERP adapter: Helios/Pohoda/SQL first, SAP as optional plugin

Hardware Profiles for Clients

Profile	Hardware	PoC Price	Annual License
Profile A	Blackwell cluster	€20-25K	€8-12K
Profile B	RTX 4090	€12-18K	€5-8K

Model governance: €2,000-3,500/year recurring.

Latency SLAs

Path	Target
Rule-based known fault → work order	<5 seconds
LLM analysis + SQL inventory	<60 seconds
SAP ERP write	<5 minutes
HITL notification delivery	<30 seconds

Tech Stack

Layer	Component
Telemetry ingestion	MQTT/OPC-UA → Mosquitto
Anomaly detection	Threshold rules + LLM semantic log analysis
Agent orchestration	n8n or custom OpenClaw agent
ERP connector	Modular REST adapter (Helios/Pohoda/SQL, SAP plugin)
HITL approval	Mattermost + approval webhook
Monitoring	Prometheus + Grafana (ops signals only, no production data)
Inference	Local Blackwell cluster (Nemotron-Super-120B via vLLM)

Phase 0 — Deployed (2026-03-20)

- IOL Agent stack running on **qa-docker** (10.1.2.99)
- Containers: iol-db (PostgreSQL 16), iol-agent, iol-bot — all on Docker
- Dashboard: http://10.1.2.99:8080/
- Features: Equipment view, alerts, work orders (approve/reject), inventory, agent audit trail, fire-demo button
- Migrated from Podman to Docker (2026-03-20)

Project Phases

Phase	Scope	Timeline
Phase 0 ✅	Foundation — environment, data pipeline, simulator	Week 1-2
Phase 1	Decision A — work order automation (MVP)	Week 3-6
Phase 2	Decision B — autonomous procurement	Week 7-10
Phase 3	Supply chain optimization	Week 11-14
Phase 4	Commercialization — packaged offering	Week 15-16

Data Sovereignty

- All production data stays on-premise (sensor readings, maintenance logs, ERP data)
- Monitoring collects only operational signals (CPU/GPU/latency/decision counters)
- Remote access optional via Cloudflare tunnel to Grafana only
- IATF 16949 compatible: air-gapped deployment supported

Open Questions (Awaiting Caly)

#	Question	Impact
1	Simulated vs real sensors for PoC?	Phase 0 scope
2	Target ERP for Phase 1? (mock/Helios/Pohoda/SQL)	2-day vs 2-week scope difference
3	Grant/funding application project?	Deliverables structure
4	IATF/ISO compliance requirements?	Architecture lock
5	Existing prospect or spec work?	Urgency and billing
6	Sample maintenance logs available?	LLM analysis quality

---

4. Key Milestones (March 2026)

Date	Event
2026-03-10	Token burn incident ($78) — subagents routed to Claude API instead of local LLMs
2026-03-12	TRT-LLM compatibility proxy built, local subagents operational
2026-03-16	Cloudflare migration planned (9 services), Caddy cleanup (19 stale routes)
2026-03-17	Local inference subagents fully operational (3 endpoints, all 4 agents tested)
2026-03-18	GPU cluster final deployment: Ollama production-ready, RDMA verified, NCCL + OpenMPI compiled
2026-03-19	Nemotron-3-Super-120B deployed on dual-node vLLM (Ray), all subagent routing confirmed
2026-03-19	IOL Agent project briefed, full team planning session, Phase 0 deployed to qa-docker
2026-03-20	Podman → Docker migration on qa-docker, IOL dashboard live

---

5. Known Issues & TODO

- [ ] Clean up stale TRT-LLM proxy services (18355/18356/18350)
- [ ] Tensor parallelism (tp=2) blocked by MPI BML initialization — vLLM workaround in place
- [ ] Set SMARTEMAILING_WEBHOOK_SECRET on stage environment
- [ ] Harry CRM: fix `NEXT_PUBLIC_API_URL` (baked at build time), fix `_get_targeted_customers()`, rebuild stage Docker image
- [ ] Harry CRM handover documentation for Michal Ruják (Fincross)
- [ ] IOL: answer open questions to unblock Phase 1

Generated by Koda 🐾 — LMP Consulting AI Agent Ecosystem