From Pilot to Scale:

Deploying Sovereign AI in Manufacturing

Benefits, Barriers and Real World Impact

Gorka Unamuno
ADRA | Multiverse Computing

RUDOLFOZTF VO 2026
Novo mesto, 31st March 2026

MULTIVERSE COMPUTING

The AI Data Robotics Association

Why this matters now

Industrial AI is moving from prototype to operational capability

RU DOL FO VO ZTF 2026

Novo mesto, 31st March 2026

Industrial AI is entering a new phase

- The technology is maturing
- Industrial pressure to adopt is increasing
- The real challenge is no longer only experimentation

Where AI is already proving value

Where AI is already proving value

Optimisation

Scheduling, process tuning, energy and resource allocation

Quality inspection

Machine vision, anomaly detection, defect classification

Predictive maintenance

Condition monitoring, failure prediction, maintenance planning

Decision support

AI assistants for operators, engineers and supervisors

Why many companies still struggle to scale

The real barriers are usually not where people expect

AI usually does not fail in the demo.
It fails in deployment.

- Clean demo environment
- Limited scope
- Manual support in the background
- No real integration burden

What usually blocks industrial AI adoption

- Fragmented and inaccessible data
- OT/IT integration with legacy systems
- Cybersecurity and confidentiality requirements
- Compute and energy constraints
- Governance, trust and compliance

The deployment question

The key question is not only “which model?”
It is also “where should it run?”

- Cloud
- On-premise
- Edge
- Hybrid

Different environments require different architectures

Architecture	Best fit	Main challenge
Cloud	Centralised services, scalable training	latency, data exposure, dependency
On-premise	Control, compliance, sensitive operations	infrastructure complexity
Edge	Low latency, resilience, local decisions	hardware constraints
Hybrid	Best balance in many cases	orchestration complexity

Why efficiency matters

In manufacturing, model efficiency is strategic

- Smaller models
- Lower energy consumption
- Lower hardware requirements
- Faster inference
- Better fit for edge and on-prem deployments

Why sovereign AI matters in manufacturing

Sovereign AI in industry means control

RU DOL FO VO ZTF 2026

Novo mesto, 31st March 2026

- Control over data
- Control over infrastructure
- Control over model access and updates
- Control over compliance and operational risk

Compression and practical deployment

Smaller and more efficient models can unlock adoption

- Enable deployment on constrained hardware
- Reduce infrastructure cost
- Reduce energy demand
- Improve responsiveness
- Support more sovereign architectures

Lessons from real deployments

What successful deployments usually have in common

- Start from a real operational pain point
- Design for deployment from the beginning
- Work with existing systems, not ideal ones
- Treat cybersecurity and compliance as design inputs
- Measure operational value, not only model accuracy

Three illustrative examples

Three practical deployment patterns

- **AI for quality inspection**
  Vision-based defect detection close to the line
- **AI for maintenance and troubleshooting**
  Technical knowledge support for operators and technicians
- **AI for optimisation and local decision-making**
  Fast responses under operational constraints

How to evaluate impact properly

How should companies measure impact?

- **Operational KPIs**
  downtime, scrap, throughput, lead time
- **Technical KPIs**
  latency, robustness, uptime, accuracy
- **Economic KPIs**
  energy, infrastructure cost, total cost to operate
- **Adoption KPIs**
  usage, trust, repeatability, integration effort

A practical roadmap for manufacturers

A pragmatic path from pilot to scale

- Prioritise a high-value use case
- Assess data and integration readiness early
- Choose the right deployment architecture
- Build security, governance and ownership into the design
- Scale only what proves value in operations

Final takeway

industrial AI is real

RU DOL FO ZTF VO 2026

Novo mesto, 31st March 2026

- This is not mainly a model problem. It is a deployment problem.
- In manufacturing, AI must fit the reality of operations.
- The question is not only what AI can do, but where it should run and under which constraints.
- Sovereign AI is, in practice, about control over critical assets and risks.
- Efficiency is becoming a strategic condition for adoption.
- A successful pilot is not the same thing as a scalable capability.
- Operational value matters more than technical elegance.

Ilustrative examples

Next-Gen Customer Support Chatbots

Serving 8,000 customer service agents with 50% lower costs and zero accuracy loss

Llama 3.1 8B

Industry: Telecommunications

Client: Telcomm. provider with global presence

Goal: Compress LLM to use it within an internal RAG chatbot for agents in client's stores.

Reduced costs

associated with the use of expensive API services (GPT, Gemini, etc.)

Speedup

Reduced latency – faster responses

Efficiency

Low power consumption

Security

Increased data privacy

Accelerating Legal Insights in Energy & Mobility with GenAI

Helping automate consultation on decarbonization & sustainability law analysis through model compression

Llama 3.1 8B

Industry: Energy & Mobility

Client: Top-tier Energy & Mobility operator and IBEX 35 member

Goal: Development of a Gen AI system for legal literature consultation.

Reduced costs

associated with storage requirements and energy consumption

Speedup

Tested on advanced infrastructure (8 A100 Tensor Core GPUs)

Efficient

Lower consumption compared to LLMs

Accurate

Precise and relevant responses

Empowering Tax Authority with an Intelligent IRPF Virtual Assistant

Delivering faster, more consistent tax guidance for regional administration through trusted AI automation

RU DOL FO VO ZTF 2026
Novo mesto, 31st March 2026

Industry: Public Sector / Taxes

Client: Regional public tax authority in Spain responsible for managing and enforcing personal income tax (IRPF) processes across the territory.

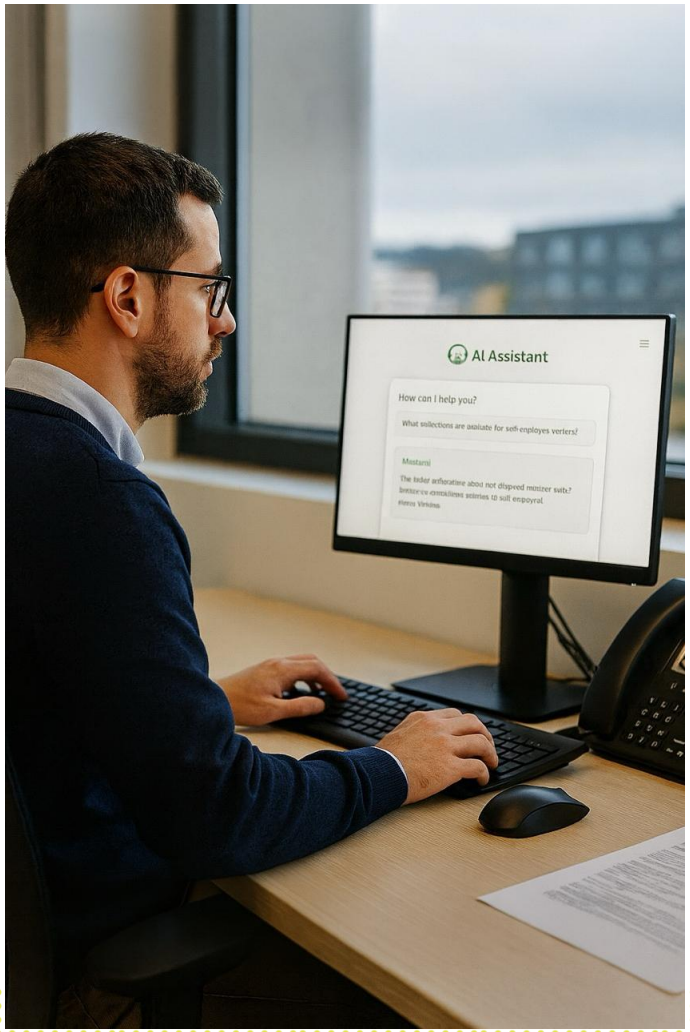
The Challenge:

Tax experts handle large volumes of complex IRPF regulations and case-specific inquiries. **They needed a secure, reliable, AI-powered virtual assistant** to support decision-making, improve efficiency, and reduce the time spent searching, interpreting, and cross-checking tax information.

The Solution:

We developed a **Generative-AI IRPF Virtual Assistant** designed to support tax specialists with accurate, contextual and compliant guidance:

- **Custom LLM trained and adapted** to IRPF documentation, workflows and expert-level reasoning.
- **RAG-based architecture** retrieving and grounding answers in Hacienda's validated tax knowledge.
- **Secure, GDPR-compliant deployment** integrated with client's infrastructure, ensuring full confidentiality and controlled access.



Faster expert guidance on IRPF rules, cases and interpretations

Higher consistency and accuracy in answers supported by authoritative documentation

Reduced operational workload for tax specialists through automated information retrieval

Real Time Border Surveillance for Defense & Military

Mission Critical Object Detection using Satellite Images Processing of 670K km^2

RU DOL FO VO ZTF 2026
Novo mesto, 31st March 2026

YOLO v8-x

Industry: Defense

Client: Transnational defense and security organization

The Challenge:

This Defense and security organization needed to perform real-time object detection over massive areas using high-resolution, multi-spectral satellite imagery. The key challenge was **to scale inference speed and refresh rates** while controlling infrastructure and energy costs, **without compromising accuracy** in mission-critical operations.-

The Solution:

Multiverse Computing compressed the YOLOv8-x model using CompactifAI, delivering a version that is **much smaller, faster, and more efficient, with no loss in accuracy**. This enables scalable, high-frequency satellite image processing with lower latency, reduced power consumption, and significantly lower operational costs, making real-time border surveillance viable at scale. The project validated that our technology makes it possible to run advanced AI models in highly constrained environments such as satellites, enabling real-time, edge-based border surveillance without reliance on large-scale infrastructure.

Example image

Enhancing Pedestrian Detection with Compressed Vision Models

Helping a leading automotive manufacturer improve road safety through on-edge compressed AI compression

RUDOLFOVO ZTF 2026
Novo mesto, 31st March 2026

YOLO v8-m

Industry: Automotive

Client: European supplier of advanced electronic systems for automotive

The Challenge:

The customer needed to significantly improve pedestrian detection accuracy using RGB + LWIR data, while **deploying the model on embedded automotive hardware with strict latency, power, and compute constraints**. Existing state-of-the-art models delivered strong accuracy but were too slow and energy-intensive for real-time, on-edge deployment.

The Solution:

Multiverse compressed the YOLO-based vision model using CompactifAI technology, achieving **a much smaller and faster model with minimal impact on detection quality**. The resulting solution delivers real-time inference, lower energy consumption, and improved suitability for embedded automotive systems, enabling safer pedestrian detection without hardware redesign.

Example image

Enhancing In-Car Voice Assistants with Compressed AI Speech Models

Providing a leading European automaker with 30% smaller models delivering faster, smarter, higher-quality voice AI

RUDOLFOVO ZTF 2026
Novo mesto, 31st March 2026

Style TTS

Industry: Automotive

Client: Leading European automotive manufacturer

The Challenge:

The automaker needed to deploy **a high-quality in-car voice assistant on constrained automotive hardware,** requiring significant model compression without degrading the naturalness, clarity, or intelligibility of the assistant's voice. Existing speech models delivered strong audio quality but consumed too much memory and compute for large-scale, in-vehicle deployment.

Our Solution:

Using CompactifAI, Multiverse compressed the Style TTS model by over 30%, reducing both parameters and memory footprint **while maintaining — and in some metrics improving — audio quality**. The compressed model enables faster, more efficient on-device voice AI, improving scalability and user experience without compromising sound quality.

Efficient Object Detection with Compressed Vision Models

Deploying low-power YOLO models on FPGA devices for real-time ship detection

RU DOL FO VO ZTF 2026

Novo mesto, 31st March 2026

YOLO v8-n

Industry: Earth Observation

Client: Consortium of Spanish aerospace and technology companies

The Challenge:

Deploying real-time object detection from stratospheric platforms required extremely low power and memory consumption, while still maintaining reliable detection accuracy. Standard YOLO models were too large and resource-intensive to fit into FPGA-based, on-device environments, limiting real-time performance.

Example image

The Solution:

We applied progressive model compression and quantization to YOLOv8-Nano, fitting the full architecture into dedicated FPGA memory. Two compression levels were evaluated: Soft compression, achieving moderate size reduction with near-original accuracy, and Strong compression, maximizing memory and power savings with only marginal accuracy degradation. **This enabled low-power, real-time ship detection fully on-device, without reliance on cloud processing.**

Empowering Technical Support in Manufacturing with Compressed LLMs

Helping a global industrial technology leader extract key insights from user manuals through efficient on-device GenAI

Llama 3.1 8B

Industry: Industrial Technology

Client: Leading multinational in industrial technology & manufacturing

Goal: Extract key information on technical manuals to help technical support queries using RAG technology.

Operation:

Upload technical documents. After a query, the system returns the answer found in each of the documents and returns the parts of the document used to create the answer.

Example:

Implementation into a mobile device to make queries without internet access.

Compression

Reduce the model size from 8B to 3.2B parameters

Optimization

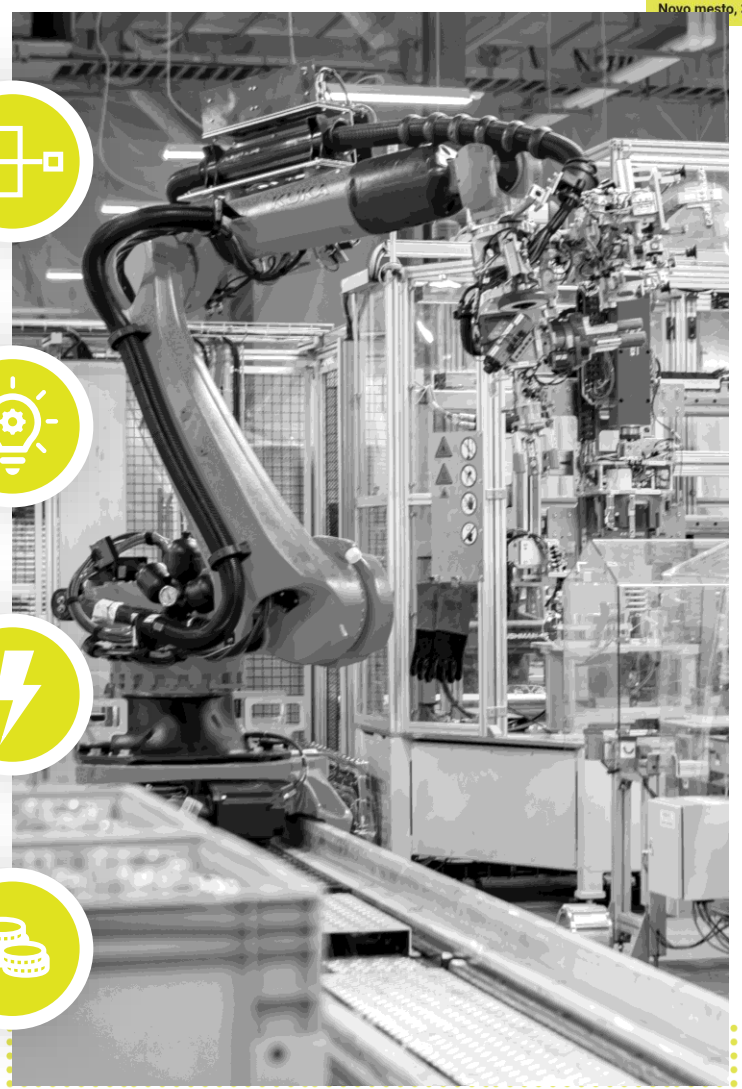
Reduce computational requirements, enabling implementation for embedded systems.

Speedup

Increased inference speed

Reduction

of the Operational costs



Compressed AI Model for Zero-Defect Manufacturing

Reducing production defects by 6.2× while improving efficiency and training speed

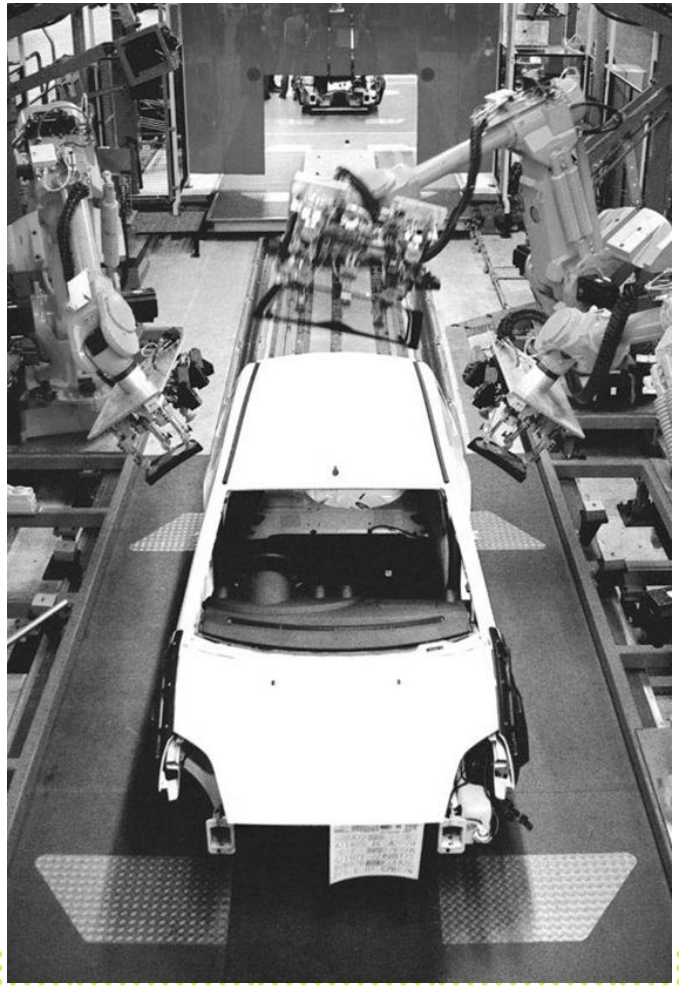
RU DOL FO VO ZTF 2026

Novo mesto, 31st March 2026

VGG16

Industry: Manufacturing

Client: Multinational leader in advanced manufacturing and industrial automation



The Challenge:

Zero-defect manufacturing requires extremely high detection accuracy while operating at industrial scale and speed. Traditional deep learning models were **computationally heavy, slow to train, and costly to deploy**, limiting their practicality for real-time quality inspection across large production lines.

The Solution:

We applied CompactifAI compression technology to a VGG16-based vision model, achieving a 4.6× reduction in parameters while preserving virtually identical accuracy (~0% F1 loss). This efficiency gain translated into **a 6.2× improvement in defect detection, up to 16% faster training, and significantly lower computational costs**, making high-precision, scalable quality inspection viable in real-world industrial environments. The resulting model was designed and validated for deployment on edge and industrial devices, enabling real-time defect detection close to the production line.

Quantum ML Quality Control for Automotive Components

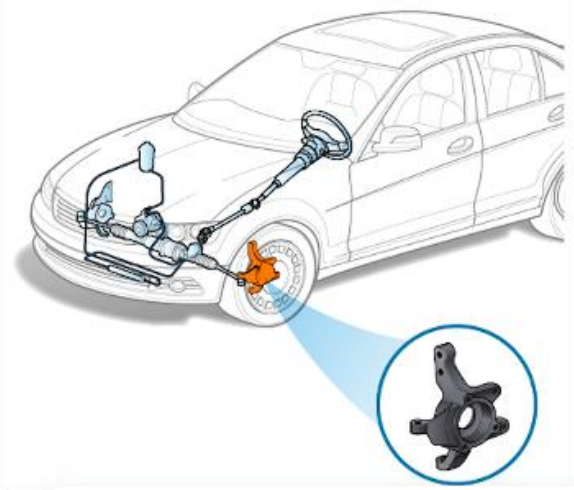
Leveraging quantum-inspired Deep-learning computer vision model to spot invisible defects and prevent costly rework

RU DOL FO VO ZTF 2026
Novo mesto, 31st March 2026

Industry: Manufacturing - Automotive

Client: Applied research institute driving innovation in industrial automation and computer vision

Use a **quantum-inspired deep-learning computer vision model** to assess images in real time and identify defects which may be undetectable for the human eye.

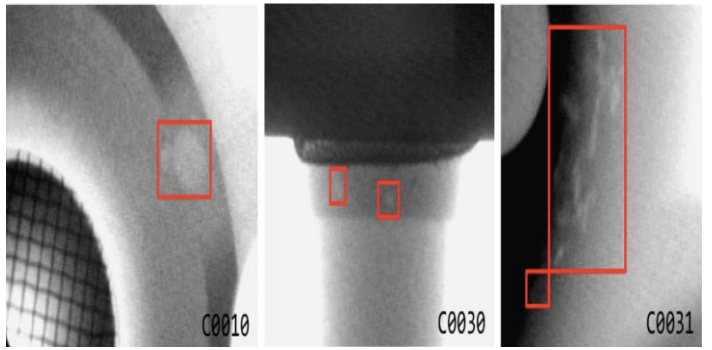


The Problem:

- In the **Die and Cast Process** for car parts rotator
- Use computer vision to detect defects that the **human eye can´t see**
- **Wrong tool and die cast can be very costly** so computer vision helps them detect defects.

The Solution:

- Quantum Inspired
- Uses **custom trained images** to identify what is considered good and defect products.
- **Explains why it is defective** and highlights the area.



Accelerating Clutch Configuration Optimization

Solving complex, large-scale optimization problems to achieve 1000× faster computation

Industry: Manufacturing
Use Case: Clutch Configuration Optimization
Client: Automotive component manufacturer

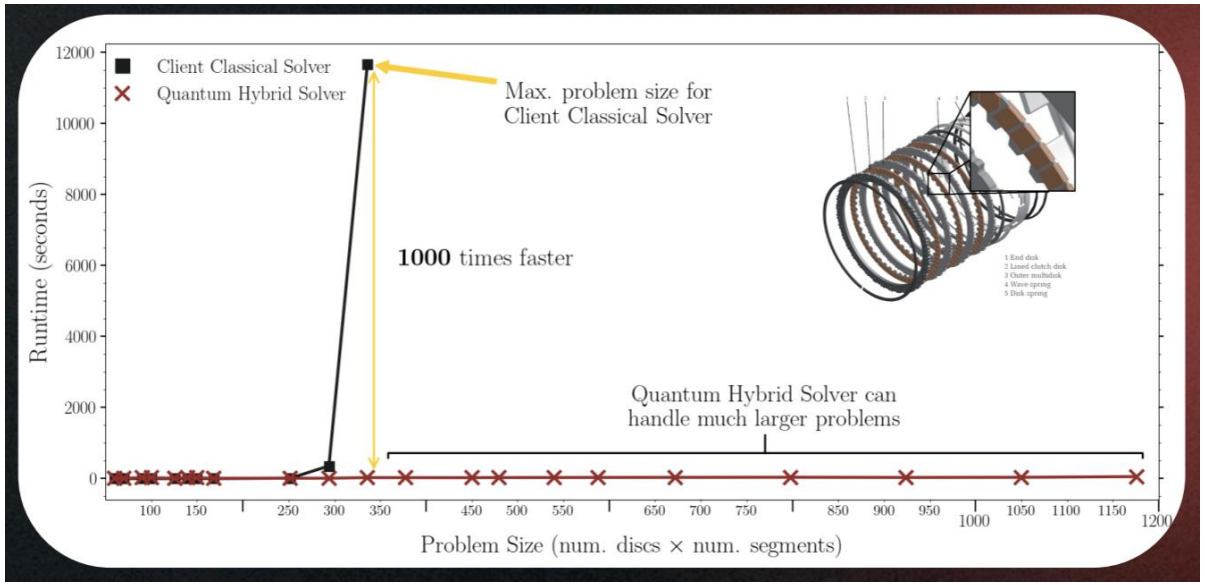
Optimization of clutch configurations during manufacturing is critical to ensure and improve the functioning and durability of engines.

Difficult optimization problem:

- Large number of clutch configurations.
- Problem quickly becomes **intractable for classical methods**.

Our quantum solution:

- Handles largest classically-solvable problem **1000x faster**.
- Solves problems **3x larger** with little increase in runtime.



AI-Powered Simulation for Faster Design Validation Using Graph-NN

Accelerating part design and cutting computational simulation time through geometric machine learning models

Industry: Manufacturing

Client: Global leader in metal forming solutions

The Challenge:

Currently, formability simulation (used to determine whether a part can be stamped without defects) requires **costly and time-consuming finite element analysis**. Each design iteration involves several simulations before reaching a viable geometry.

The Solution:

Accelerate the design and validation processes of stamped metal parts using **graph-based geometric machine learning models** capable of quickly predicting material thinning during the stamping process.

The Results:

The model drastically reduces stamped part validation time by avoiding multiple full simulations, providing a fast and sufficiently accurate estimate of material behaviour. Additionally, by relying on 3D geometry represented as a graph, it captures complex spatial dependencies that other models fail to represent.

Faster Design Cycles

Shorter development time for new designs

Enhanced Efficiency

Increased productivity for engineering teams

Lower Costs

Reduced simulation and hardware costs

Seamless Integration

Potential for integration into a smart CAD–CAE workflow

Optimizing Energy Storage

Helping Iberdrola improve grid efficiency with quantum & quantum-inspired optimization

RU DOL FO ZTF VO 2026
Novo mesto, 31st March 2026

Industry: Energy
Use Case: Network Battery Placement Optimization
Client: Multinational energy provider

The Challenge:

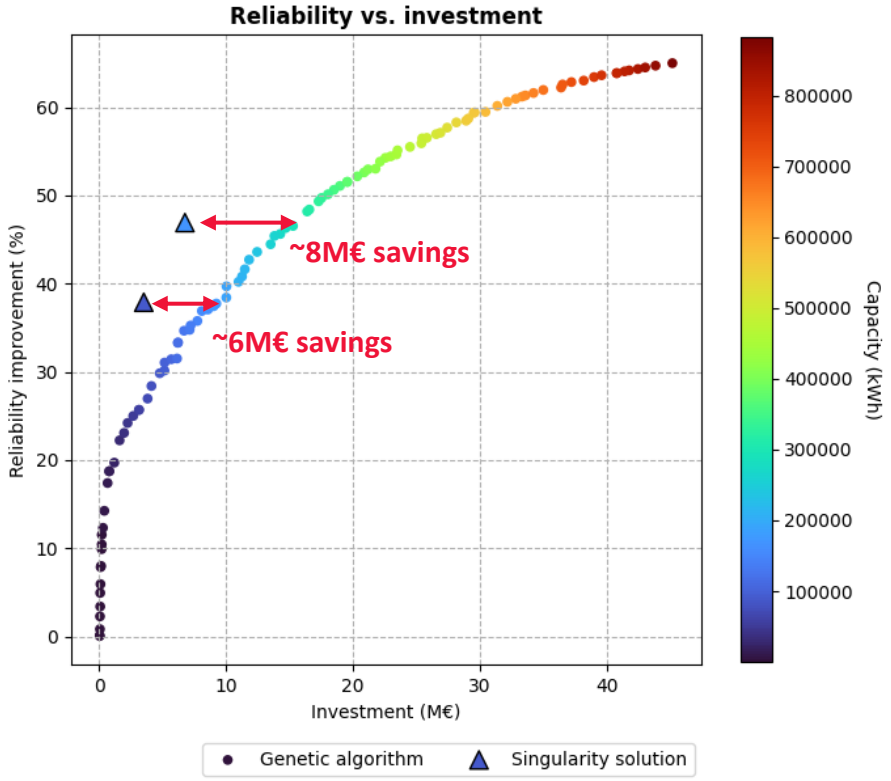
As power grids incorporate growing volumes of renewables, electric vehicles, heat pumps, and other distributed energy resources, **battery storage has become essential to maintain system balance.**

Without optimized battery placement and sizing, **grids risk facing instability, higher operational costs, energy losses, and underutilization of renewable generation**—ultimately undermining the reliability and sustainability of the entire network.

The Solution:

Singularity Optimization found optimal battery placements that would obtain the same reliability as the client solution with investment savings of millions of euros.

- **Investment:** Optimizing the cost of purchasing and installing multiple batteries within the power grid.
- **Voltage control:** Maximizing the grid's ability to maintain voltage levels across network nodes.
- **Reliability:** Minimizing the impact of power outages on end customers.



Optimized Placement

Quantum-inspired algorithms identified the most efficient locations for grid-scale batteries, enhancing stability while avoiding unnecessary installations.

Reduced Costs

The optimization model minimized both investment and operational costs by improving asset utilization and reducing redundant deployments across the grid.

Improved Efficiency

Enhanced voltage control and energy distribution improved overall grid performance and renewable integration under real operating conditions.

Empowering Industrial Maintenance with Multimodal AI Assistants

Helping a top-tier energy provider enhance field operations with next-generation RAG technology

Industry: Energy

Client: IBEX 35 global energy leader with over 100M customers worldwide

The Challenge:

Energy companies manage vast amounts of technical documentation (manuals, schematics, and reports containing text, images, and tables) that technicians need to consult during maintenance tasks at plants.

Finding specific information in real time is complex and time-consuming, increasing operational costs and slowing incident resolution.

The Solution:

A multimodal Retrieval-Augmented Generation (RAG) system that:

- **Analyzes and vectorizes documents, images, and tables**, storing them in a vector database.
- **Uses multimodal models (MLLMs)** to understand queries combining text and image.
- **Enables interaction via text, image, or voice,** adapting to different environments (e.g., smartphone or AR glasses).

1. **Multimodal RAG** with text and image
2. Integration of **voice interaction**
3. **Information retrieval directly from voice,** without converting it to text (via semantic voice embeddings)

Instant Access
to critical information during maintenance operations

Reduced Times
Shortened search and incident resolution times

Improved safety and reliability
through accurate queries based on official documentation

Enabling Real-Time Multimodal AI for Defense Robotics

Deploying advanced vision-language reasoning on resource-constrained through extreme model compression

RUDOLFOVO ZTF 2026
Novo mesto, 31st March 2026

Industry: Defense

Client: National defense R&D organization, focused on advancing technologies for military, security and autonomous systems

Used model: **Molmo 72B VLM**

The Challenge:

Deploy a 72B-parameter multimodal Vision-Language Model (VLM) on resource-constrained robotic platforms for real-time reasoning, while meeting **strict requirements in memory, compute, energy efficiency and data privacy**, and preserving the model's multimodal accuracy for mission-critical use cases.

The Solution:

Multiverse Computing delivered a **compressed, quantum-inspired optimized version of the Molmo 72B Vision-Language Model**, tailored for real-time deployment on edge robotic systems.

- **Quantum-inspired compression of a 72B multimodal VLM,** reducing memory and compute requirements while preserving multimodal reasoning accuracy.
- **Hardware-aware optimization for edge robotic platforms,** enabling real-time vision–language inference under tight resource constraints.
- **Delivery of a general compressed model ready for secure fine-tuning with private defense data,** fully aligned with confidentiality requirements.

On-edge multimodal reasoning

Advanced vision–language AI enabled directly on robotic platforms

Reliable & secure

Designed for privacy-sensitive and defense-grade environments

Reusable architecture

Adaptable across different robotic and autonomous hardware systems

Enabling Computer Vision on Tactical Drones Through Compressed AI

Boosting onboard autonomy & mission performance by deploying EdgeTAM on defense-grade processors

RU DOL FO VO ZTF 2026
Novo mesto, 31st March 2026

EdgeTAM

Industry: Defense

Client: Tech. organization developing advanced UAV & robotic solutions

The Challenge:

Deploy high-accuracy segmentation and object tracking on tactical drones operating with restricted compute, memory, and power budgets.

The Solution:

Multiverse used CompactifAI to compress and optimize the EdgeTAM model for onboard deployment.

Profiling

Measure RAM, latency, throughput. Identify key blocks

Compression

↓**35%** in Spatial Perceiver
↓**25%** in Decoder block

Deliverables

ONNX & TensorRT Models for In-Drone deployment

No reliance on ground compute

Autonomous capabilities directly on the drone

Reduced latency & energy consumption

Optimizes the bottleneck module (Spatial Perceiver)

More onboard AI functions per device

Multiple AI functions in parallel within the same hardware budget

Accelerating Nuclear Fusion Simulations with Compressed Diffusion Models

Improving time-to-insight in time-sensitive fusion workflows by optimizing small diffusion models for faster inference

RUDOLFOVO ZTF 2026
Novo mesto, 31st March 2026

Industry: Energy / Nuclear Fusion

Client: A leading European engineering and technology firm in the energy sector

The Challenge:

The customer needed to run **diffusion-based simulation models under strict runtime constraints**, where inference speed directly impacts the efficiency of nuclear fusion analysis workflows. Their models were already relatively small (100M and 10M parameters), but still required further optimization to reduce latency while maintaining the quality needed for reliable simulation outputs.

The Solution:

We applied CompactifAI compression techniques to two customer-developed diffusion models, **optimizing architecture and execution efficiency to accelerate inference without compromising output quality**. The work focused on block removal strategies and healing techniques to preserve model performance while achieving significant speedups for simulation pipelines.

Faster time-to-insight for **time-sensitive simulations**

Nearly 2× faster inference without sacrificing output quality

Production-ready acceleration for industrial simulation pipelines

[1] In PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index), higher is better. | [2] In RSME (Root Mean Squared Error), lower is better. | [3] In Pearson correlation, closer to 1 is better

Powering On-Device AI at Scale with Compressed LLMs

Enabling high-performance, local AI assistants on leading manufacturer's AI PCs through model compression

RU DOL FO VO ZTF 2026
Novo mesto, 31st March 2026

Industry: Electronics & Device Manufacturing

Client: Global technology leader in personal computing, printing and edge AI solutions

Used model: GPT-OSS 20B

The Challenge:

Deploy large language models locally on next-generation AI PCs while meeting **strict performance, latency, memory and quality constraints** across multiple hardware configurations (16GB, 32GB and 64GB), without degrading safety, reasoning or user experience.

The Solution:

Multiverse Computing delivered a **portfolio of compressed large language models** tailored for the client's AI PC product lines, ensuring high performance under tight memory and latency constraints.

- **Compressed versions of GPT-OSS 20B** optimized for 16GB, 32GB and 64GB AI PC SKUs
- **Performance-driven optimization** meeting strict throughput, latency, RAM and quality benchmarks defined by the client
- **Production-ready delivery with continuous model updates,** aligned with future releases of the base open model

True-on-device AI at scale
Enables local AI assistants without cloud dependency

Working in low memory
Optimized for low RAM usage with strong throughput and latency

Enterprise-grade quality
Less than 2-5% quality degradation vs original models

Future-proof model updates
Alignment with new releases of the base open models

Mass-market readiness
Designed for deployment across millions of consumer AI PCs

Najlepša hvala