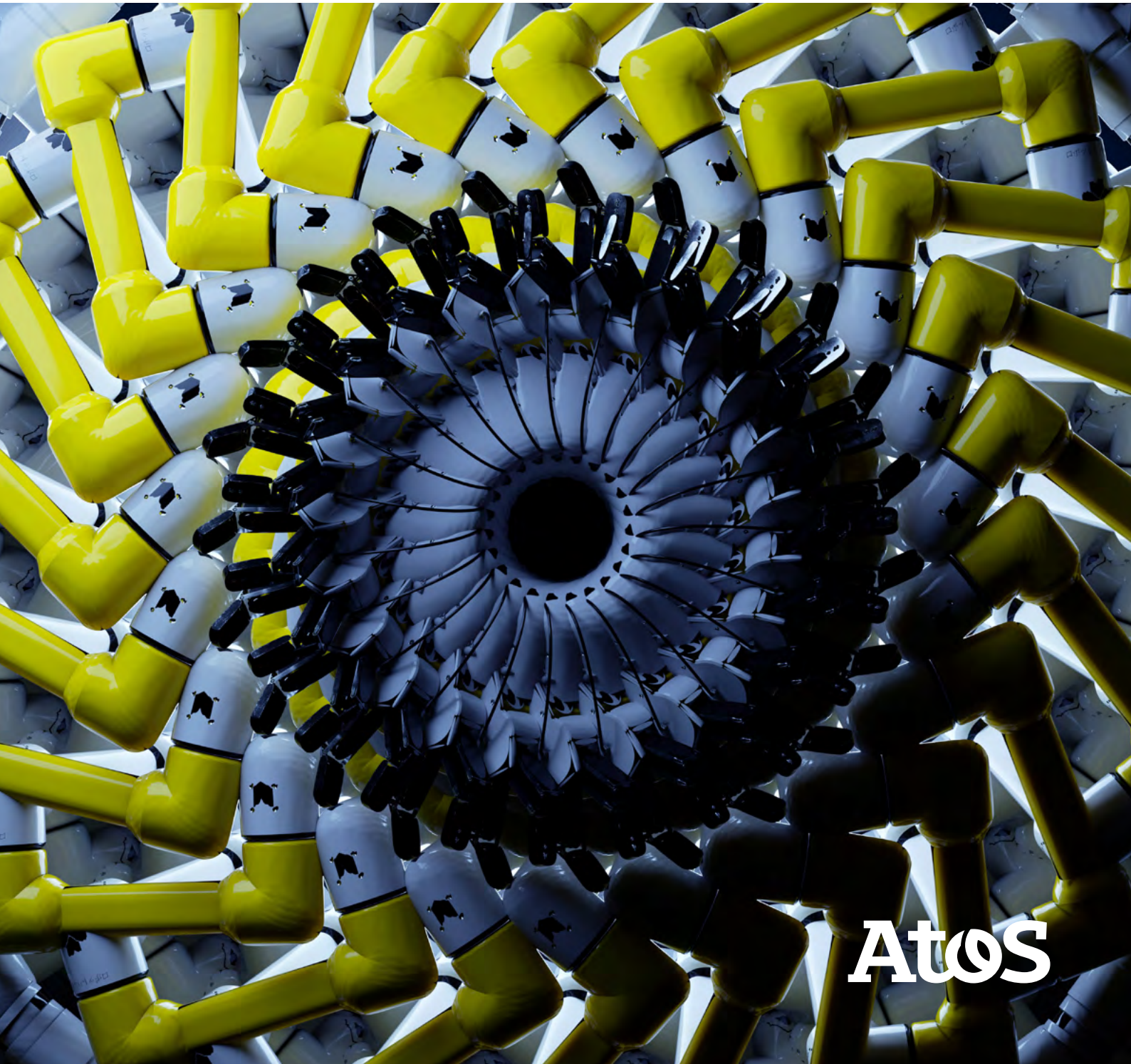
Future Makers
Research Community

Sovereign Generative AI for Manufacturing

Engineering Design Interpretation at Scale



Contents

Executive Summary

Artificial Intelligence (AI) has long been transforming the manufacturing industry through applications such as computer vision, digital twins, robotics, and predictive maintenance.

While these technologies have enhanced automation and efficiency, the emergence of Generative AI (GenAI) introduces a new paradigm: the ability to interpret, reason, and interact with complex, human-centric information such as engineering drawings and specifications.

On the other hand, in the face of engineering talent shortages, increasing compliance demands, and mounting cost pressures, discrete manufacturers are actively seeking new levers for productivity. This white paper presents a practical application of sovereign, on-premises GenAI technology to automate and accelerate engineering design interpretation tasks.

Conducted as a real-world proof of concept (PoC), the solution delivered over 2,000 hours “business impact and ROI potential” of engineering time savings by interpreting complex design standards, validating schematic drawings, and reducing manual rework. The approach offers a scalable model for engineering transformation, particularly in environments with strong Sovereign requirements.

AI in Manufacturing: Understanding the landscape

Industry context: The productivity imperative

Global discrete manufacturers face an acute challenge: engineering demand is rising while the talent pool shrinks. A Rockwell Automation study reports[1] that 83% of industrial firms aim to use AI to close this skills gap. When talking to the Manufacturing departments in Asia, design and compliance processes—especially those involving legacy documents and schematic interpretation—are prime candidates for intelligent automation.

Traditional CAD tools and document repositories provide limited support for understanding and applying design standards. The result is a heavy reliance on experienced engineers, manual rule interpretation, and costly design validation cycles.

[1] rockwellautomation.com/en-au/capabilities/digital-transformation/state-of-smart-manufacturing.html

Problem statement

When talking to multiple manufacturers, a common pattern was identified:

- Manual standards interpretation: Engineers spent hundreds of hours reviewing compliance standards (e.g., ISO, ANSI) for schematics.
- Inconsistent rule application: Subjective interpretation led to errors and rework.
- Knowledge drain: Critical knowledge resided with seniors.
- On premises availability: Many productive tools were available as SaaS service but due to IP constraints, the use of the same is limited.

The PoC aimed to determine if sovereign, on-prem GenAI could do the following:

- Parse unstructured engineering standards and drawings
- Automatically apply rules and extract tolerances
- Generate actionable design insights and annotations
- Read and interpret engineering schematics, understanding geometry, tolerances, annotations, material specifications, and assembly relationships.
- Assist engineers interactively, answering complex queries like:

  "What is the maximum tolerance specified for part X?"

  "Compare the hollow shaft keyway drawing against actual machined output."

AI and Sovereignty in Manufacturing

As the manufacturing industry rapidly advances toward full digitalization, Artificial Intelligence (AI) has become a crucial enabler, especially in areas like computer vision, digital twins, and robotic automation.

Among these, a new frontier has emerged: the use of GenAI to understand, interpret, and reason over complex engineering drawings — a task traditionally reserved for highly skilled engineers.

What is Sovereign AI?

Sovereign AI refers to AI solutions that are hosted, operated, and managed entirely within an organization's own IT infrastructure — ensuring all data, models, and outputs remain fully under the company's control.

In industries like manufacturing, aerospace, defense, and automotive, where data security, intellectual property (IP) protection, and regulatory compliance are critical, sovereign AI is emerging as a key strategic enabler.

AI is everywhere

79%
of corporate strategies **see** AI and Analytics **as critical to their success** over the next 2 years[1]

33%
of organizations are **using GenAI regularly** in at least one business function[2]

70%
of organizations will **have operationalized AI architectures** by 2025[3]

But...

48%
only of AI projects **make it** from pilot **to production**[4]

39%
of functional leaders **consider Data Protection and Privacy as a challenge** with the implementation of GenAI[5]

[1] Gartner Newsroom, July 2023
[2] McKinsey, The State of AI, Aug. 2023
[3] Gartner, Hype Cycle for AI, 2021
[4] Gartner, How Generative AI Puts Organizational AI Maturity to the Test, April 2024
[5] Gartner, Generative AI 2024 Planning Survey

Sovereign AI provides several critical advantages:

Data Sovereignty

All engineering drawings, specifications, and derived insights are retained within the company's own trusted IT environment — not exposed to SaaS Providers.

IP Protection and Competitive Advantage

Companies often guard their most valuable innovations — whether patented or held as trade secrets — as a source of competitive differentiation. Sovereign AI solutions ensure sensitive intellectual property remains fully protected

On-premises Deployment

The AI solution is deployed inside the organization's owned or directly controlled data centres — not simply at a third-party provider nearby. Full physical and logical control remains with the company.

Customizability

AI models and workflows can be adapted to meet unique industry needs, proprietary processes, or national regulatory standards.

Compliance with Legal Frameworks

Organizations can meet strict regulatory requirements (e.g., GDPR, ITAR and national data sovereignty laws) without depending on the legal interpretations of external vendors.

Cost Control and Predictability

Operating sovereign AI solutions allows companies to predict and manage the total cost of ownership (TCO) without being subjected to unexpected cloud usage charges or dependency risks.

Manufacturers operate in an intensely competitive environment where design data and engineering specifications are strategic assets. Losing control over such data — whether due to breaches, vendor dependencies, or regulatory non-compliance — can threaten a company's competitive position.

While modern cloud services offer sophisticated security mechanisms, the reality is that companies often seek greater autonomy, transparency, and control over their critical engineering data and AI workflows. Sovereign AI fulfills this need by ensuring full ownership and governance of both the data and the models, without compromising operational efficiency.

Proof of Concept: AI-powered Drawing Interpretation

In this section, we outline the Proof of Concept (PoC) developed to evaluate the feasibility and impact of using GenAI for interpreting complex engineering schematic drawings. We begin with a real-world use case that highlights the challenges faced by engineers without AI assistance. From there, we explore the AI-driven solution architecture, examine how the engineer experience changes with AI support, and conclude with a summary of the measurable business impact and return on investment (ROI). This section lays the foundation for understanding how AI can move from experimentation to value delivery in a real production environment.

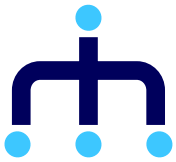
Demystifying the tech: Key terms you should know

This section introduces key technical terms that appear after this section. It's written to help non-technical readers understand the technologies behind the Proof of Concept (PoC) and how they influence performance, accuracy, and decision-making. Familiarity with these concepts will help everyone stay aligned as we move from experimentation toward implementation.

Key technical terms explained:

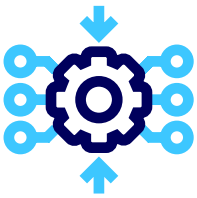
Large Language Model (LLM)

A powerful AI model trained on vast amounts of text data to understand, reason, and generate human-like language. LLMs can answer questions, interpret documents, and even understand drawings when combined with other tools.



Fine-tuning

The process of adapting a general LLM to a specific use case (like engineering or manufacturing) using domain-specific examples. This improves accuracy for specialized tasks.



Inference

The stage where an AI model makes predictions or generates output based on input data. In simple terms, it's when the trained model is "put to use."

Training

The process of teaching an AI model by feeding it data so it can learn patterns, relationships, and context. Training is computationally intensive and usually done before deployment.

Retrieval-augmented Generation (RAG)

A method that helps an LLM give better answers by retrieving relevant information from external sources (e.g., databases or documents) and combining it with its built-in knowledge. This is helpful when the model needs up-to-date or domain-specific data.

Containers

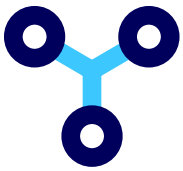
Lightweight, portable packages that include everything needed to run software—code, libraries, settings. Containers (like Docker) make AI systems easier to deploy consistently across different environments.

Function (in LLM context)

A callable operation or tool that the LLM can invoke—such as searching a document, summarizing text, or extracting dimensions from a drawing. Functions extend the model's utility.

Schematic Drawing

A technical diagram that shows the structure, components, and dimensions of an object or system. Interpreting these drawings accurately is crucial for engineering applications.



Multimodal Learning

The AI's ability to process and connect multiple types of input—like images and text—simultaneously. This is key to interpreting engineering drawings alongside their written specifications.

Real-world use case: Interpreting schematic drawings

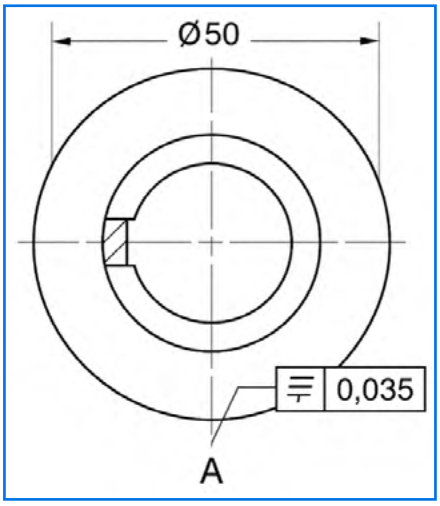
As part of the PoC, we explored a range of schematic drawings, from simple to highly complex, to assess the models' ability to accurately interpret engineering diagrams and extract key design parameters. These included geometric features, dimensions, tolerances, and annotations.

We aimed to ensure the solution could not only read and understand the technical drawings but also interpret associated tables and correlate them with descriptive text. This correlation was tested across multilingual scenarios, particularly in English and Mandarin, to evaluate the system's robustness in international engineering contexts.

To validate the PoC's performance, we engaged multiple engineers to independently analyse the same set of drawings. Their findings were compared against the model's output to measure accuracy, consistency, and relevance.

Below is one of the sample drawings used during the evaluation:

A drawing shows a hollow shaft (Ø50 mm OD, Ø30 mm ID) with a 6 mm wide, 3 mm deep internal keyway. The symmetry tolerance is 0.035 mm relative to Datum A. Symmetry of the keyway is critical for correct key fitting and load transmission, controlled by a symmetry tolerance of 0.035 mm to the central axis as per ISO standards. The shaft is press-fitted inside a cylindrical housing, providing radial and axial stability during torque transmission.



Diameter Range (D)	Aspect Ratio (D/L)	Symmetry Tolerance (mm)
≤ 8 mm	D/L ≥ 0.75	0.035
> 8 mm and ≤ 22 mm	D/L ≥ 0.75	0.035
> 22 mm	D/L ≥ 0.75	0.040 or per design spec
All	D/L < 0.75	0.035 + adjustment (per ISO)

Engineering without AI: Current pain points

Challenge	Description
Time	Reading a schematic took 20–60 minutes depending on complexity
Error risk	Frequent mistakes interpreting tolerances, datums, and dimensions
Skills gap	High training needs to maintain precision
Throughput	Engineering capacity limited to X parts/day
Traceability	Interpretations vary across engineers
Multilingual	Need to support both English and Chinese

AI-enabled solution architecture

To empower engineers without compromising data security or sovereignty, our Proof of Concept (PoC) deployed a fully sovereign AI stack within a private hosted environment. This architecture supports advanced GenAI capabilities while keeping sensitive engineering data like CAD files and specifications, with no reliance on external cloud providers. More details on the architecture is as detailed below.

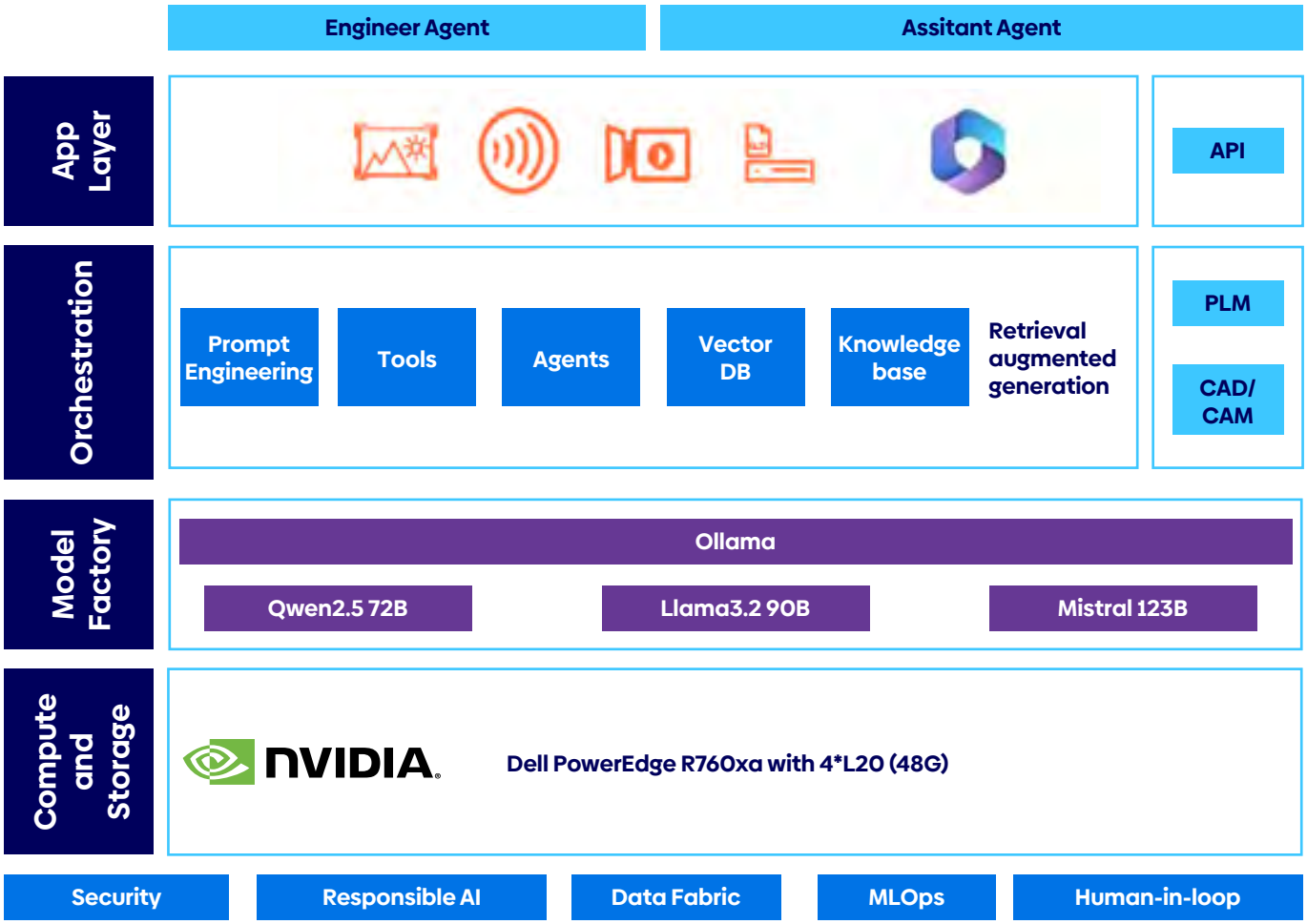
While hardware was hosted in the customer's secure data center, our emphasis was first on selecting the right AI software architecture and components, which were later powered by high-performance NVIDIA compute resources.

Enhanced engineer experience with AI

- Engineer uploads drawing or 3D model.
- AI extracts features, tolerances, annotations via Retrieval-Augmented Generation (RAG).
- Standards are matched and applied automatically.
- Results (tables/comments) guide quick engineering decisions.
- Engineer adjusts design based on validated feedback.

Technology stack and LLM evaluation

Below is the high-level technical architecture deployed for the PoC. The architecture was designed to support the ingestion, interpretation, and analysis of complex engineering drawings and documents through a combination of modern AI techniques and scalable infrastructure.



The core components are detailed in the following sub chapters:

	What it is	What it does	In the process
Compute and Storage Layer	This is the infrastructure foundation that powers the entire solution	• The system runs on a Dell PowerEdge R760xa server equipped with 4 NVIDIA L20 GPUs (48GB VRAM total). • This setup enables a fast and secure on-premises model inference, essential for protecting sensitive IP like engineering schematics. • It ensures low-latency processing, critical for interactive design review and compliance queries.	It provides the raw horsepower needed to run large models like Alibaba Qwen, Meta LAMA, and Mistral within the enterprise firewall, preserving sovereignty and performance.

Model Factory Layer	The LLM runtime environment responsible for executing large language models	• Hosted on Ollama, which simplifies model deployment and switching between models.
		• Supports multiple foundational models:
		  • Qwen2.5 72B for Chinese and contextual understanding
		  • LLaMA3.2 90B for English and structured data processing
		  • Mistral 123B for tasks involving visual reasoning and design interpretation
			This layer handles the heavy lifting of natural language and visual interpretation, enabling engineers to ask complex design questions in natural language.

Orchestration Layer	The middleware brain that manages how inputs are processed and passed through the AI stack	Combines several subsystems: • Prompt engineering: Optimizes how queries are phrased for the models • Tools and agents: Logic-based workflows such as compliance checker and geometry analyzer • Vector DB: For semantic similarity search on diagrams, past designs, or regulations • Knowledge base: Repository of internal policies, ISO/ASME standards and tribal knowledge • Retrieval Augmented Generation (RAG): Brings external documents like ISO PDFs, internal standards into the model's context for precise answers.	This layer enables contextual grounding, so the AI doesn't hallucinate. It ensures queries are accurate, relevant, and regulation-aware.

Application Layer	The user interaction and task automation interface	• Contains Engineer Agent (user input) and Assistant Agent (AI response logic).
		• Supports:
		  • Image/diagram input
		  • CAD file parsing
		• Connects to external systems via API, PLM, and CAD/CAM integrations.
			This is where engineers upload files, ask questions, and receive structured outputs like tables, flagged violations and suggestions. It brings the AI's capabilities into existing tools.

Governance and Enablement Layer	The operational backbone that ensures reliability, safety and integration	• Security: Ensures design IP and AI outputs are protected • Responsible AI: Enforces bias detection, fairness, and audit logging • Data fabric: Manages how data flows between systems and tools • MLOps: Supports model updates, monitoring, and performance tuning • Human-in-the-loop: Allows engineers to override AI or validate results; critical in regulated environments	This layer ensures the system is enterprise-grade, secure, and compliant, and can be improved over time without downtime.

LLM evaluation results

During the PoC, we evaluated various LLMs to determine their effectiveness in understanding technical engineering content, including both simple and highly complex schematics.

Small Large Models (SLMs)

Both small-scale models with fewer parameters were tested. While the smaller models offered faster inference and lighter resource consumption, they lacked the precision and contextual depth required for interpreting detailed engineering drawings.

Huge Large Models with billons of parameters

In certain complex scenarios, we considered increasing the model size or parameter count to improve performance. However, the results showed diminishing returns—accuracy did not significantly improve despite increased computational cost.

After trialing a variety of models across different configurations and scenarios, we arrived at a balance between performance, accuracy, and resource efficiency. Based on this, we shortlisted three LLMs that demonstrated the most promise for interpreting complex engineering design documents while maintaining practical inference times and integration flexibility.

Model	Strengths	Performance	Limitations
Qwen 2.5 72B	Excellent for Chinese language processing and customer-specific scenarios.	High accuracy and efficient resource usage	Primarily suited for simpler tasks and Chinese language contexts
Llama 3.2 90B	Excels in the English language, understanding and data preprocessing tasks	Great at handling complex queries in English but struggles with certain complex Chinese queries	Less effective for non-English languages, particularly Chinese
Mistral 123B	Demonstrated strong performance in analyzing diagrams and computational tasks, critical for analyzing design blueprints. This model demonstrates superior output presentation capabilities. For example, as per the client's request, it can present information in a table format.	Stable and consistent, well-suited for high-performance computational tasks in manufacturing workflows.	Can require significant computational resources for large-scale deployment.

Fine-tuning path

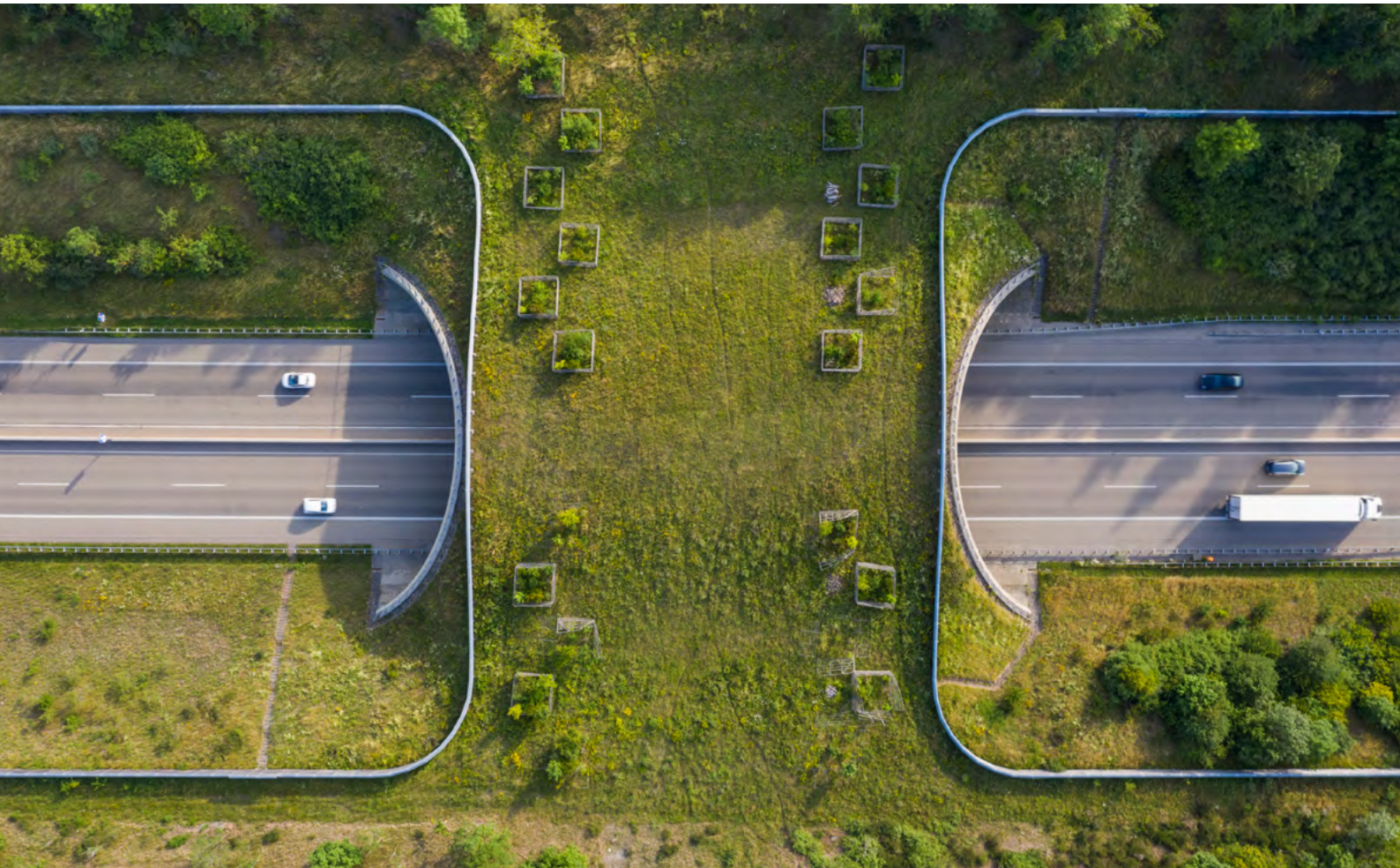
As part of the forward roadmap, fine-tuning LLMs on engineering-specific data has been identified as a promising approach to boost accuracy and contextual understanding, especially for interpreting tolerance notations, dimensions, and domain-specific annotations.

Risk and mitigation strategies

While sovereign AI offers significant benefits such as data sovereignty, IP protection, and AI alignment with local regulations, it is essential to address the potential challenges and actively plan mitigations to ensure a successful implementation.

Below are the key risks and strategies to address them:

Risk	Description	Mitigation Strategy
High initial investment	Sovereign AI requires substantial upfront costs for infrastructure and expertise	• Start with public cloud experimentation ("in the wild") to gain experience, check the model and the performance, for security perspective on public cloud, leverage synthetic data
Integration complexity	Legacy systems like PLM/CAD/CAM are difficult to integrate with modern AI systems.	• Begin with a “design assistant” as a standalone aid • Use API-first to support phased integration
Data ingestion and preprocessing	RAG systems need to ingest large volumes of structured and unstructured data, leading to latency and quality issues	• Build scalable data pipelines • Define evaluation metrics with business teams early
Data governance	Lack of governance causes bias, stale results, and compliance issues	• Apply strong data access controls and freshness checks
Scalability and infrastructure	Under-provisioning Sovereign AI compute/storage can result in delays and cost overruns	• Size infrastructure based on actual model demands • Experiment on public cloud with synthetic data
Business Misalignment	Poorly defined KPIs or unclear use cases can lead to failed adoption	• Align AI outcomes with measurable business success • Incorporate human-in-the-loop reviews with domain experts



Business impact and ROI potential

In this section, we quantify the tangible benefits observed during the PoC. By comparing key operational metrics before and after introducing the AI-assisted workflow, we highlight measurable improvements in efficiency, productivity, and resource utilization. These outcomes serve as a strong indicator of the ROI and help justify scaling the solution across broader use cases.

Metric	Before	After
Average time per drawing	30 min	5 min
Engineering hours	2,500 hrs/year	500 hrs/year
Productivity	Manual pace - ~5,000 drawings/year	~6,000 drawings/year (1.2 x more drawings)
Training needs	Ongoing	Prompt templates

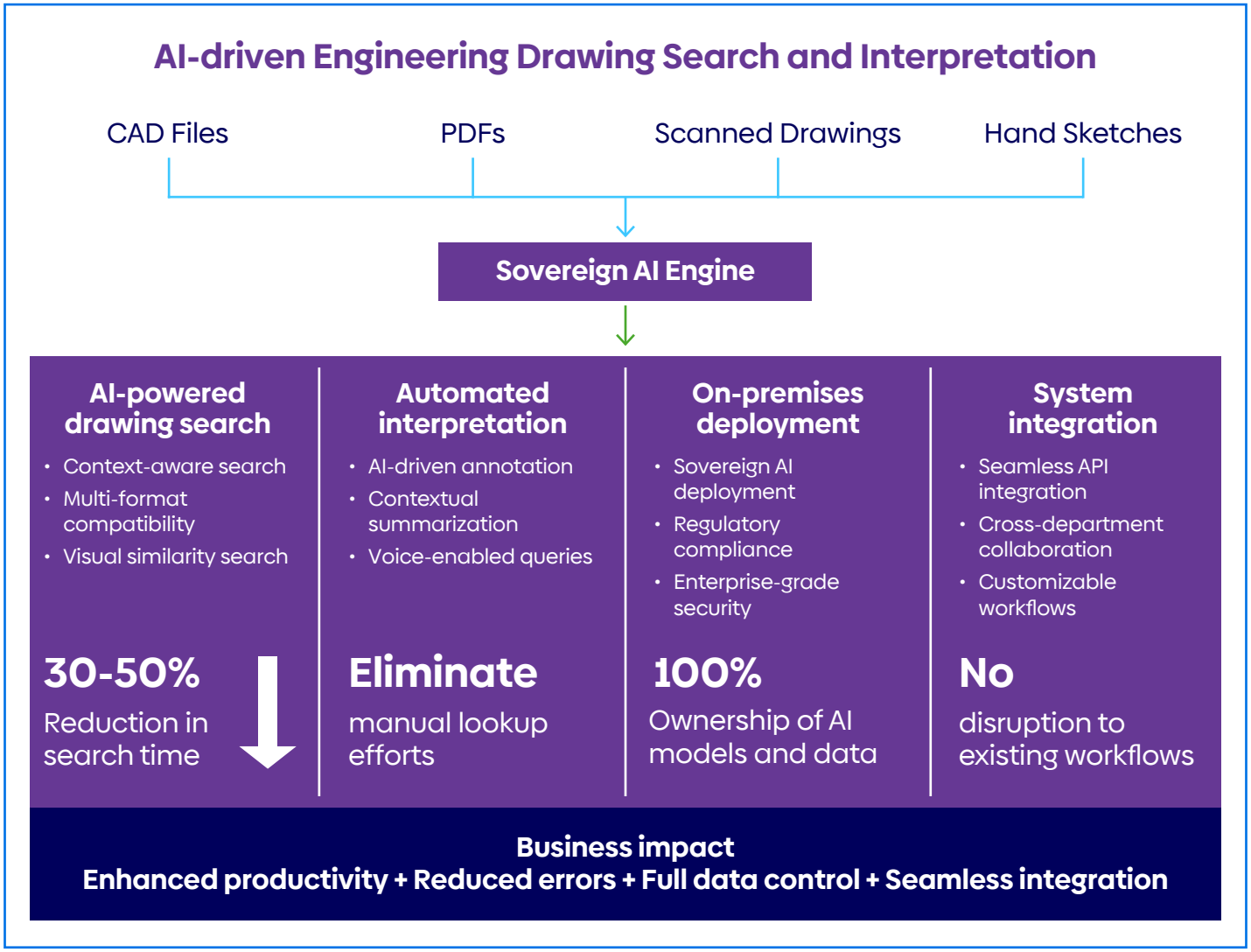
Results

- Engineering hours: Reduced from 2,500 to 500 hours/year, saving ~2,000 hours annually — equivalent to 1 full-time engineer (FTE)
- Productivity: Teams can now process 6× more drawings daily, aligned with time saved per drawing
- Improved cross-team standardization
- Faster onboarding and knowledge transfer

Conclusion and next steps

Extending beyond the POC: Path to scalable implementation

Following the successful completion of the PoC, we evaluated the technical feasibility and business impact of scaling this AI-driven approach across the enterprise. This section outlines key solution capabilities, enablers, and expansion strategies, positioning the PoC as a foundation for transformational change in engineering operations.



AI-powered drawing search

Capabilities	Business Impact
• **Context-aware retrieval:** AI models interpret engineering-specific language, symbols, and annotations to improve search relevance. • **Multi-format compatibility:** Supports diverse content types — PDFs, CAD files, scanned drawings, and hand-sketched blueprints. • **Visual similarity search:** Locates comparable components or patterns from historical designs, reducing redundancy and rework.	• Achieved 30–50% reduction in engineering search time • Improved knowledge reuse accelerates design cycles and enhances productivity

Automated drawing interpretation and annotation

Capabilities	Business Impact
• **AI-driven feature extraction** Detects and classifies dimensions, material specifications, tolerances, and geometric constraints. • **Contextual summarization** Converts complex visual data into structured, machine-readable insights	• Minimizes manual interpretation efforts, boosting engineer productivity • Reduces downstream design errors and rework through consistent data extraction

Multimodal integration: Unlocking new potential

Objective: Expand AI capabilities to support multimodal inputs—integrating text, images, audio, and video into a unified interpretation layer.

Use Cases	Impact
• Upload a schematic (image) and a written note (text); the AI cross-validates design elements and regulatory compliance. • Ingest training videos or operational footage; the AI extracts key procedural steps and safety warnings. • Process audio notes from field technicians; AI transcribes and classifies actionable issues.	• Reflects real-world engineering workflows where documentation spans across multiple formats. • Enhances the AI's ability to reason across modalities, improving usability in practical environments.

Fine-tuning for domain-specific accuracy

As part of the roadmap, we explored whether expanding Large Language Model (LLM) parameters significantly improved outcomes on complex drawings. Our testing indicated diminishing returns beyond a certain point.

Refined Strategy	Next Steps
• Instead of expanding model size, we identified fine-tuning as a more optimal path for improving performance in highly technical scenarios. • Fine-tuning with proprietary engineering documents, drawing annotations, and domain-specific terminology offers a high-ROI enhancement.	• Curate a fine-tuning dataset from validated historical data and expert-reviewed interpretations. • Evaluate model improvements against defined accuracy benchmarks.

Standardizing access with Model Context Protocol (MCP)

To simplify AI integration at scale, we recommend investigating the Model Context Protocol (MCP) — an emerging standard that unifies how AI agents access and interact with external data.

Benefit:

- Enables structured access to design documents, compliance records, and production data without bespoke integration
- Functions as a universal data interface for AI—comparable to USB-C for hardware
- Reduces friction in AI deployment and maintenance across enterprise systems

Cross-industry applicability

Though this initiative focused on discrete manufacturing, the architectural and AI design patterns are readily extensible to:

Aerospace
where compliance, tolerances, and traceability are critical.

Automotive
supporting fast-paced design iteration and quality control.

Industrial equipment
for managing complex, variant-heavy product lines.

Outcome: The solution is well-positioned as a scalable, domain-agnostic AI platform adaptable to multiple sectors.

Conclusion

This initiative has demonstrated that when thoughtfully integrated with engineering data and domain- specific requirements, GenAI can significantly transform how technical teams interact with complex design documentation. The PoC validated the feasibility of AI-driven interpretation, search, and summarization across multimodal formats such as drawings, schematics, text, and scanned documents.

Through structured evaluation of multiple LLMs, we identified scalable architectures and fine-tuning strategies best suited to industrial contexts. In particular, our focus on sovereign AI, context-aware retrieval, and domain-specific enhancements ensures the solution remains compliant, secure, and operationally viable for high-value use cases.

Looking forward, the roadmap toward full-scale implementation includes deeper integration with PLM and CAD systems, the adoption of multimodal AI capabilities, and robust on-premises deployments. With clear productivity gains, improved data accessibility, and reduced error rates, this solution offers a strong RoI, positioning the organization to lead in AI-enabled engineering innovation.

This journey marks the beginning of a broader transformation — unlocking new potential in design automation, knowledge reuse, and operational excellence.

Authors and Acknowledgements

Authors

Purshottam Purswani
CTO, Atos APAC
Future Makers Research
Community member

Daisy Zhan
AI Portfolio Lead
Atos, China

FLENDER

Guo Yan Ming
IT Head
Flender, China

About Atos

Atos Group is a global leader in digital transformation with c. 72,000 employees and annual revenue of c. € 10 billion, operating in 68 countries under two brands — Atos for services and Eviden for products. European number one in cybersecurity, cloud and high-performance computing, Atos Group is committed to a secure and decarbonized future and provides tailored AI-powered, end-to-end solutions for all industries. Atos is a SE (Societas Europaea) and listed on Euronext Paris.

The purpose of Atos is to help design the future of the information space. Its expertise and services support the development of knowledge, education and research in a multicultural approach and contribute to the development of scientific and technological excellence. Across the world, the Group enables its customers and employees, and members of societies at large to live, work and develop sustainably, in a safe and secure information space.

Find out more about us
atos.net
atos.net/career

Let's start a discussion together

Future Makers Research Community

The Future Makers Research Community is a global network of our Future Makers - exponential thinkers and forward-looking technology thought leaders - across Atos.

Our Future Makers are united by profound curiosity, a strong growth mindset and a passion for shaping the future through exponential technologies applied in a deep industry context. We collaborate on thought leadership, (co)-innovation and R&D across all innovation horizons and our ambition is to elevate organizations and drive lasting impact.

In close co-creation with our clients and partners, we deliver bold ideas and industry use cases, by anticipating trends and market needs that will reshape businesses and society.

Together, we're not just imagining the future — we're building it.

Document Accessibility Tested
Atos

This document meets high standards of accessibility

Atos is a registered trademark of Atos SE. June 2025. © Copyright 2025, Atos SE. Confidential Information owned by Atos group, to be used by the recipient only. This document, or any part of it, may not be reproduced, copied, circulated and/or distributed nor quoted without prior written approval of Atos.

106440 - JS + HC

Atos