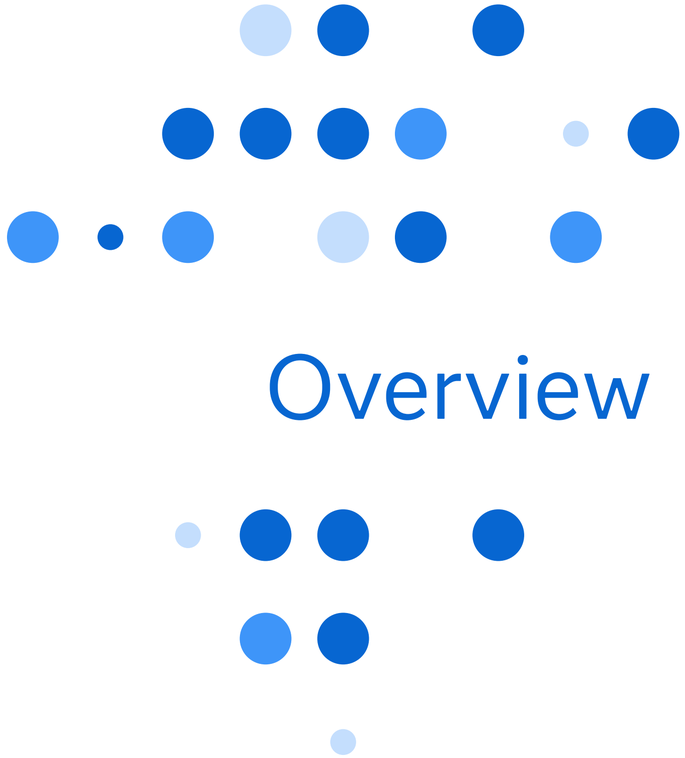
Predictive power-up

How adding a GenAI layer to predictive maintenance tools is changing what’s possible in industrial sectors

sas



Overview

01

Mission-critical in Manufacturing

The first business applications of generative AI (GenAI) have occurred in some of the most obvious (and least risky) areas. Today these tools are used extensively to assist research, streamline HR processes, enhance customer engagement in marketing contexts, and more.

For example, airlines are already using GenAI to efficiently field millions of customer queries about everything from upcoming travel plans to rewards programs and current flight changes. But would the same airline operators feel confident relying on jets that were monitored, maintained, and serviced using these technologies? They absolutely should – and they will. But the use cases for industrial sectors are more complex when it comes to machine maintenance. Plus, the stakes are higher, for obvious reasons – a poor customer service experience due to GenAI is fundamentally less risky than a machine malfunction.

Inside, we'll explore how the manufacturing industry can apply GenAI capabilities, including advanced approaches using large language models (LLMs) and retrieval-augmented generation (RAG), in one critical part of their operations: predictive maintenance. Manufacturers are already using machine learning (ML) in their maintenance operations. How can they combine GenAI advances with ML to catapult predictive maintenance strategies to new levels of productivity and performance?

02
Sound familiar?

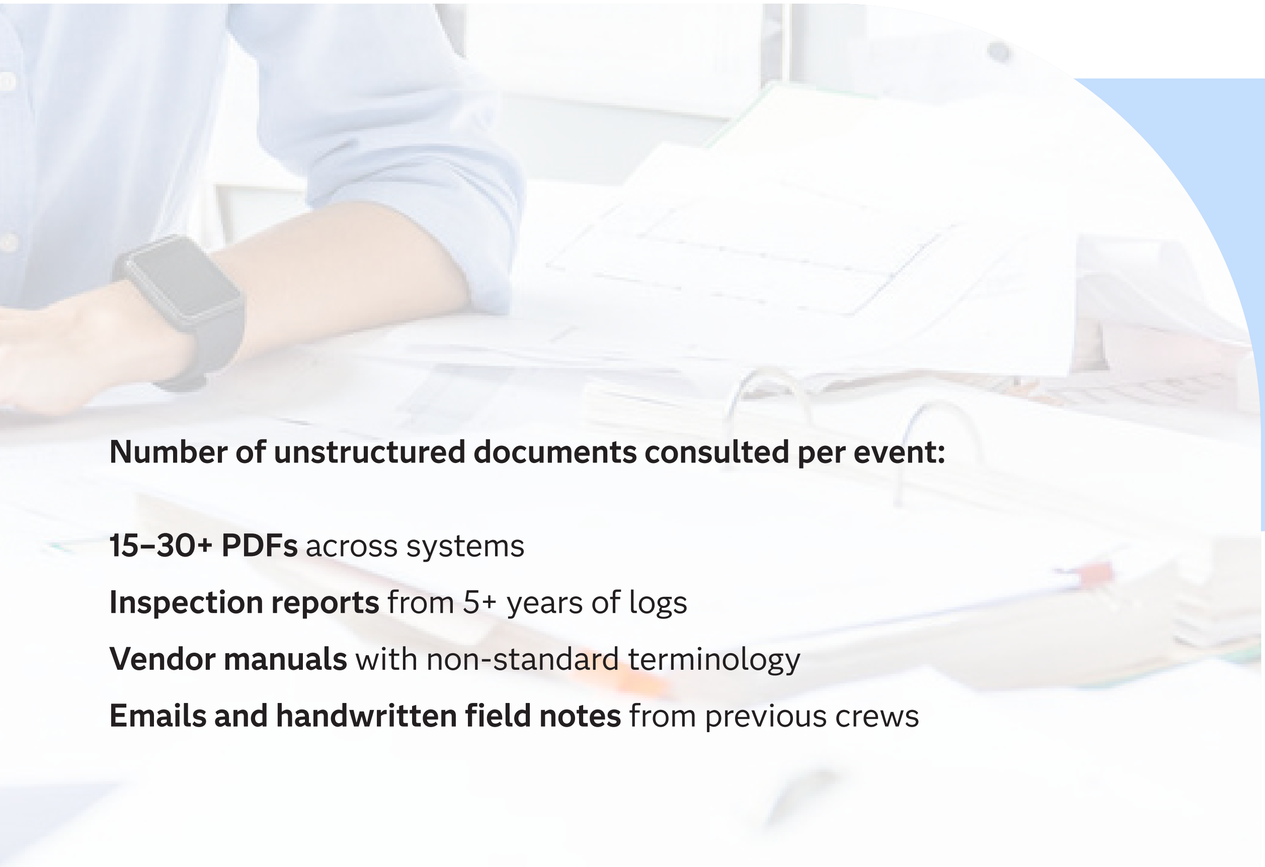
Every day, manufacturers rely on some of the most complex machinery and technology ever created, from industrial robots and CNC machines to turbines and beyond. Their complexity is what makes maintenance – preventive or otherwise – so difficult and time-consuming. Regardless of whether your organization uses gas turbines, this example reflects the layers of complexity that manufacturers face all the time.

A typical gas turbine has over 20,000 components—

each with its own maintenance procedures and dependencies.

The OEM manual for a GE Frame 9F turbine runs over 6,000 pages—

not including separate documents for the generator, auxiliary systems, or site-specific operations.



Number of unstructured documents consulted per event:

15–30+ PDFs across systems

Inspection reports from 5+ years of logs

Vendor manuals with non-standard terminology

Emails and handwritten field notes from previous crews

Field service bulletins—

can number in the hundreds per year for major OEMs like Siemens or GE, requiring constant monitoring to stay current.

Average time to identify the root cause of a turbine trip (using traditional methods)—

6–10 hours.

03

Machine learning has taken us halfway to our destination

For manufacturers, ML has already proven to be a powerful resource in predictive maintenance strategies. Using ML, manufacturers automatically monitoring large volumes of maintenance-related data – on virtually anything that can be measured.

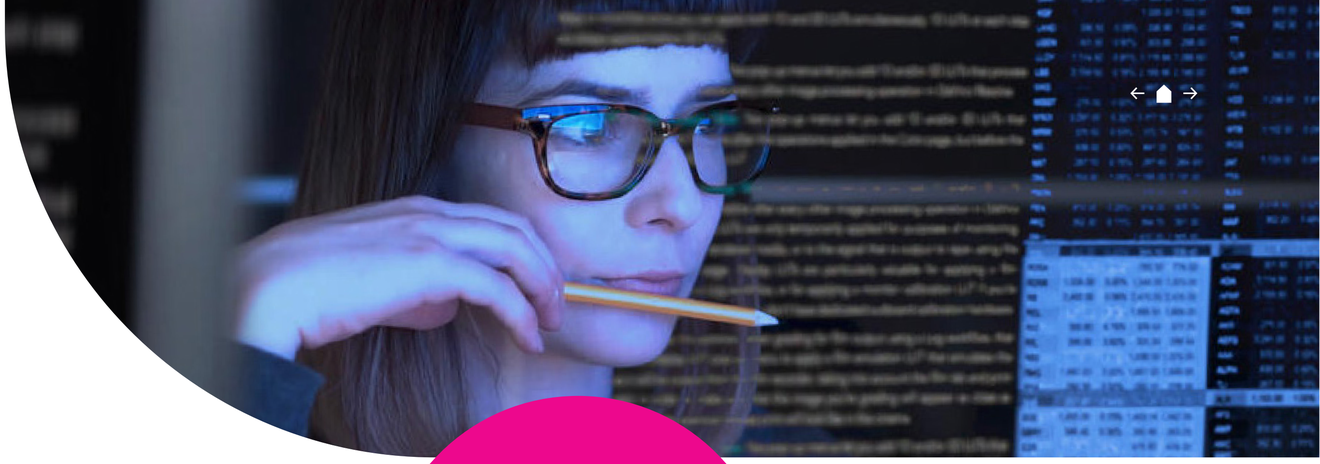
This data is used to generate outputs for maintenance organizations – lots of them. Alerts and warnings start to fill up maintenance dashboards, presenting experienced technicians with a flurry of red-exclamation-point messages so vague that each requires more investigation. *How serious is the alert? Does it require an urgent response? Does it need a response at all?*

Because these alerts and warnings are presented in a vacuum of context, they require more work, time, and expertise to unravel and act on – any alert can set maintenance professionals on a maze of SME knowledge, legacy manuals, and scattered documentation. Even the handful of those that may be legitimate can require additional expertise from a different technician with relevant skills, or from manuals or other resources.

The result: Too many alerts, and much slower decision making. That's where advanced GenAI tools such as LLMs and RAG come in.

ML-based Predictive Maintenance is powerful, but...

- Outputs often **lack context**
- Technicians get **vague alerts**
- Subject matter knowledge is **scattered**
- Decision-making is **slowed**



Warning: Combustion instability in chamber 2

ML has made it easy to detect maintenance issues like this. But detection is only part of the solution – once presented with this alert, technicians must quickly navigate a maze of subject matter expertise, legacy manuals, and documentation scattered across the organization. This is where LLMs and RAG can help.

04

What happens when you add a GenAI layer?

Your organization has already invested significant resources into developing a maintenance alert system based on ML capabilities. Do new GenAI capabilities mean you have to overhaul that system? Not at all. Those alerts just need the critical context that different information sources can provide – and GenAI is the most effective tool for tapping into those sources immediately.

GenAI capabilities – especially LLMs and RAG working in tandem, executed by AI agents under human oversight – can automate much of the time- and resource-intensive process of gathering reliable, well-documented maintenance knowledge sources. Just as important, it can serve this information to maintenance professionals in ways that make it easy to make sense of the information – and to dive deeper in situations where that is required.

05

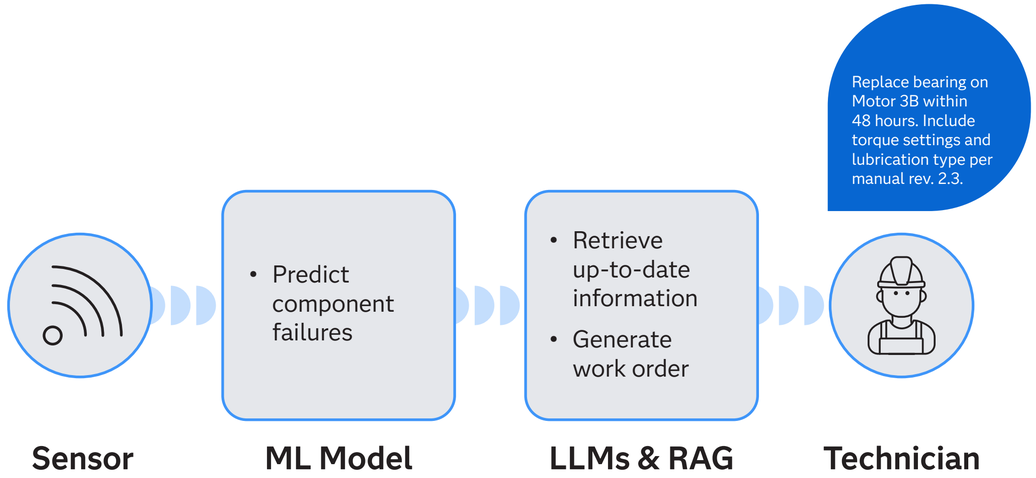
LLMs and RAG make quick work of raw source data on maintenance

For any one of the hundreds or thousands of machines operating in a manufacturing organization, there are three core data inputs that inform any maintenance intervention:

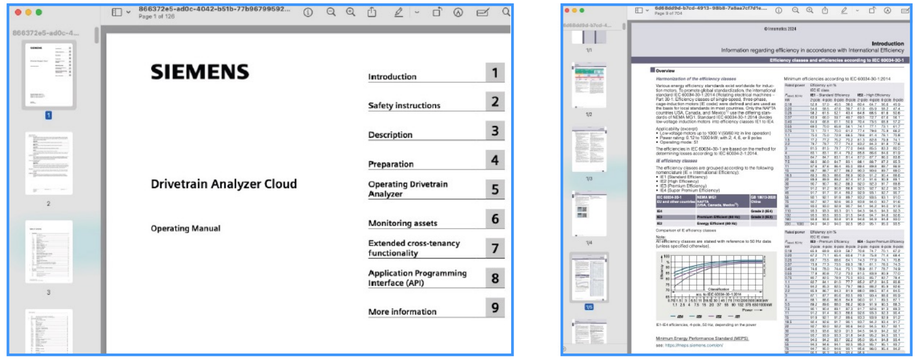
- **Current operating conditions:** How is the machine operating in real time, and are there any anomalies detected that indicate underperformance or potential failure?
- **Product information and troubleshooting guides:** If a problem is encountered, what information do we need to diagnose and solve the problem?
- **Organizational experience:** What is our experience with this machine and other identical machines?

This wide-ranging, large-volume data is typically contained in various structured and unstructured formats in different parts of the organization – real-time performance data, PDFs, vendor manuals, field service bulletins, charts, tables, emails, handwritten field notes, and more. RAG tools pull in all this data, storing it in vector databases to make it easily accessible to AI agents tasked with managing ML-generated alerts from existing maintenance systems.

In practice, this means that as soon as an alert is generated, AI agents instantly swarm on it, drawing from massive volumes of raw maintenance data to present the alert in full, well-sourced context to maintenance professionals. The resulting knowledge service generates a response at the moment it is required – and the scope of the response is limited to the range of documents.



LLMs and RAG transform ML-generated predictive maintenance alerts into more useful, manageable, and insightful cues for technicians – with reference sources.

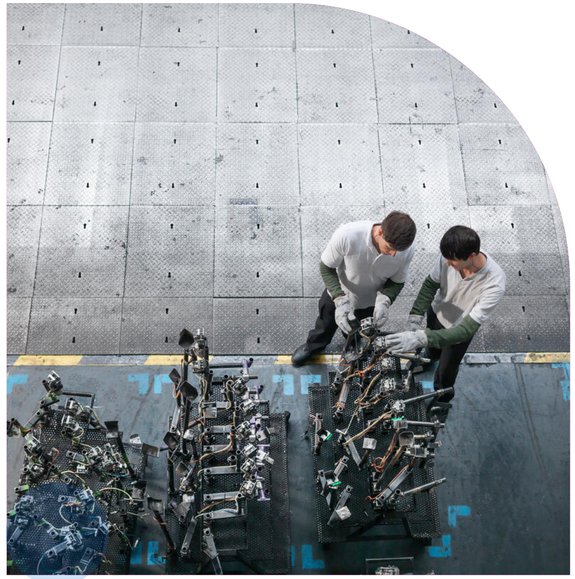


Example of a document that RAG can extract insights from - Operating manual.

06

Activate the knowledge system with AI agents

If the maintenance knowledge system (ML-driven alert systems using machine data, combined with LLM- and RAG-powered, encyclopedic information on the machines and their maintenance histories) is the beehive of maintenance data, AI agents are the worker bees that carry out all the tasks that keep the hive functioning. These agents are assigned highly specific tasks that take alerts and present them to human operators with all the relevant information available to inform decision making.



Here's how AI agents work in practice:

1

An anomaly is detected, triggering a "call" to an agent.

In SAS' approach, agents are separated from the calling applications by design – making them modular components that are reusable across a range of applications whenever and wherever they are needed. As a result, users can extract the maximum value from the time and resources required to develop these valuable agents. In deployment, API integrations bring the agents and calling applications together at runtime.

2

The agent receives basic information: "Here's the anomaly, and here's some basic information about it."

3

The agent taps into the knowledge system using targeted prompts – this is a RAG query. Through the prompts, the agent can search specific collections of information, or all of them.

"search product manuals," "search maintenance history," or "search all," for example.

4

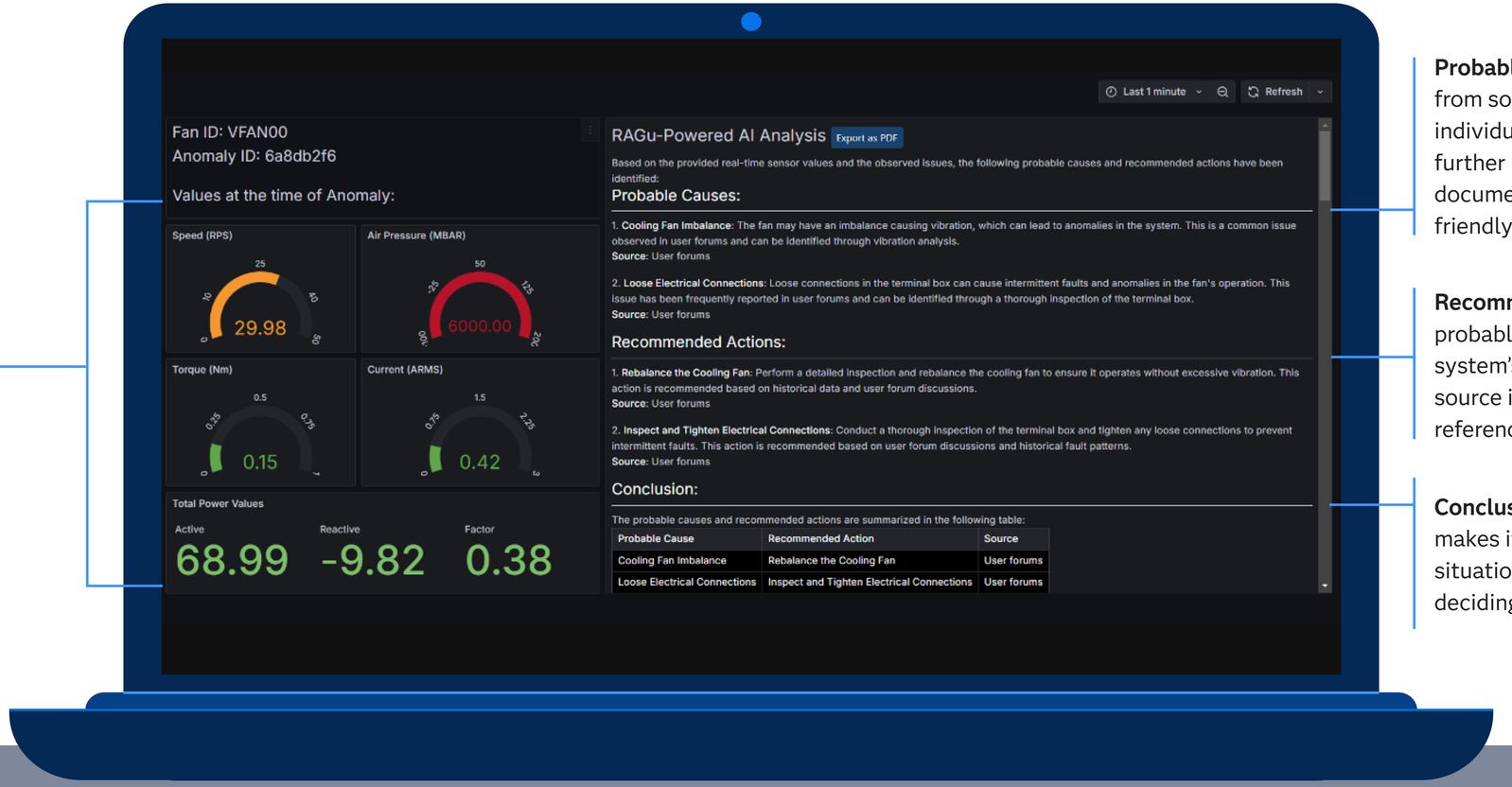
The agent queries the relevant data sources and produces a response. The response is curated in a way that is meaningful, useful, and understandable to technicians.

5

Once the agent is created, it can be redeployed as a bot into any workflow.

07
Agentic AI: What maintenance professionals see

Conditions at time of anomaly: This information is taken directly from machine monitors and performance indicators



Probable causes: These are generated from source documentation, and individual sources are referenced if further investigation is needed. Source documentation can be queried with user-friendly, AI-enabled prompts.

Recommended actions: Just as with probable causes, these actions reflect the system's "best guess" based on reliable source information, and all sources are referenced.

Conclusion: This quick-reference summary makes it easy for users to troubleshoot the situation and consider their options before deciding whether and how to act.

08

Taking the steps to scale

Large-scale agentic AI in predictive maintenance is the culmination of a series of steps. The good news is that most advanced maintenance organizations have already taken the first step: Implementing ML capabilities to detect anomalies faster.

The next step? Integrating RAG capabilities alongside existing ML tools. This can be time-consuming, mostly due to the process of identifying all the raw maintenance data required to feed into RAG models. But once that work is done, and maintenance teams begin surfacing accurate, trustworthy answers using their favorite LLMs, the momentum builds – and agentic AI moves within reach. Once maintenance organizations become adept at developing, deploying, and reusing these AI agents, all the conditions are in place for fast, efficient, accurate predictive maintenance at a scale that would have been unimaginable only a few years ago.

SAS can help. We have pioneered some of the most advanced RAG- and LLM-based predictive maintenance tools and strategies available today, and we're ready to help your organization.

For more information – and to get the conversation started – just email us.

To learn more, please visit **sas.com**

To contact your local SAS office, please visit: **sas.com/offices**