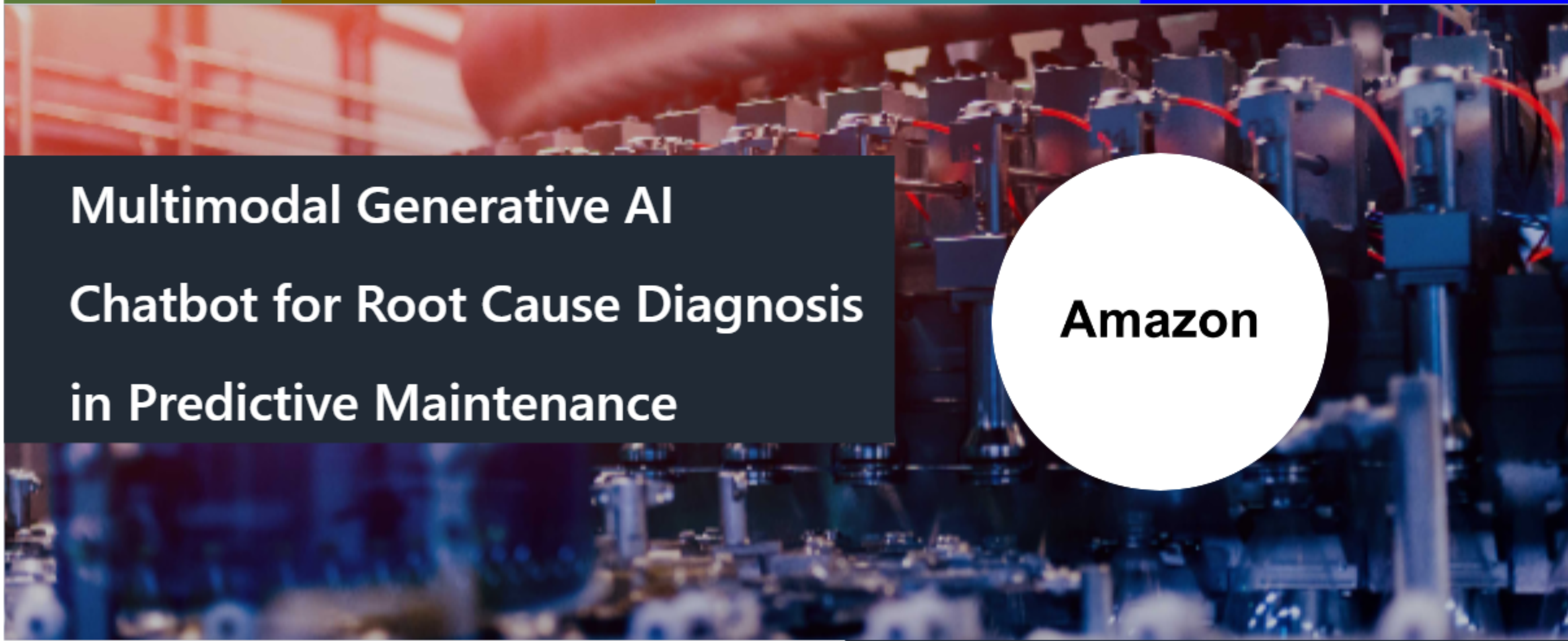
Maintenance | Industrial automation | Improved process performance | Natural language processing



Multimodal Generative AI Chatbot for Root Cause Diagnosis in Predictive Maintenance

Amazon

BUSINESS PROBLEM

Predictive maintenance systems are valuable for detecting early signs of equipment failure using sensor data such as vibration and temperature. When anomalies occur, alerts notify maintenance teams to take action before breakdowns happen. However, while these systems excel at signaling that a problem exists, they provide little support in diagnosing the underlying issue or guiding technicians through resolution. As a result, teams face delays, extended downtime, and inconsistent fixes. Technicians often rely on manual searches through complex documentation and individual expertise to identify the root cause.

DATA SOURCES

The solution uses two primary data sources: time-series sensor data capturing vibration and temperature at one-second intervals, and PDF-based repair manuals containing images, tables, and procedural text. The sensor data enables anomaly detection, while the manuals provide diagnostic context for root cause analysis.

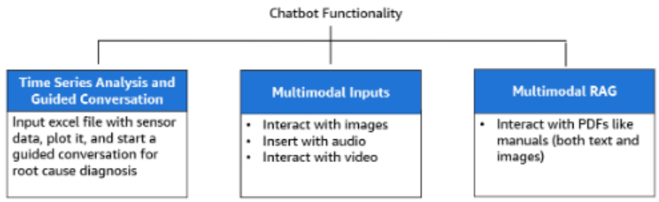
Data Types and Format

Time-series sensor data (CSV), equipment images (JPEG/PNG), video clips (MP4), audio recordings (WAV), and PDF repair manuals with structured and unstructured text.

Author: Carla Lorente Anon

APPROACH

I developed a generative AI chatbot using AWS tools to help technicians diagnose equipment issues. The assistant analyzes sensor trends, retrieves relevant repair steps, and supports image, video, and audio inputs. It uses multimodal retrieval-augmented generation (RAG) and LLMs to generate guided, real-time troubleshooting conversations tailored to the issue.



IMPACT

The AI assistant helps technicians resolve issues faster by reducing time spent searching through manuals and guessing next steps. In pilot testing, diagnostic time dropped by over 30%, saving thousands per incident in avoided downtime. It also improves the use of existing knowledge by retrieving relevant repair history and technical content. Instead of relying on static, generic work orders, technicians receive case-specific, targeted guidance. By narrowing down possible causes, the tool increases first-time fix rates, reducing repeated maintenance and improving equipment uptime. Once issues are resolved, the assistant supports better reporting by guiding technicians to log accurate closure codes. This leads to cleaner maintenance records and improves model feedback for future predictive systems. The assistant also acts as a virtual mentor for new technicians, providing consistent, guided troubleshooting that accelerates onboarding and reduces dependence on tribal knowledge. Although built for manufacturing, the architecture is flexible and can be adapted to other industries like logistics, utilities, or oil and gas—anywhere sensor data and maintenance documentation are used together.

DRIVERS

Industrial operations face rising pressure to reduce downtime and improve maintenance efficiency. While sensors can detect issues early, diagnosing them remains slow and manual. The lack of tools that connect sensor data with actionable, technician-friendly guidance created a clear need for a scalable generative AI solution that supports faster, more accurate troubleshooting.

BARRIERS

A major barrier was the difficulty in measuring long-term impact before full implementation. Metrics like reduced downtime, improved fix rates, and model retraining benefits require extended usage to quantify. Early results were promising, but demonstrating sustained value depends on continued adoption and integration into daily technician workflows.

ENABLERS

Strong collaboration with maintenance and data engineering teams enabled access to high-quality sensor data and real-world feedback. The company's openness to innovation and support from AWS resources accelerated prototyping and deployment. Cross-functional input from technicians ensured the solution addressed practical needs on the ground.

ACTIONS

I visited the company's fulfillment center to observe operations firsthand and gather insights directly from technicians. This helped identify key pain points in troubleshooting. Based on these findings, I built a generative AI assistant that analyzes sensor data, interprets technician inputs, and retrieves relevant repair guidance to support real-time diagnostics.

INNOVATION

Traditional work orders were static and generic, offering limited value during troubleshooting. This solution introduced dynamic, AI-generated guidance tailored to real-time sensor data. While retrieval augmented generation (RAG) is widely used across other company projects, this implementation was the first to integrate multimodal RAG, retrieving both text and images. The solution's novelty led to a patent filing.

IMPROVEMENT

The solution reduced diagnostic time by over 30% in pilot tests, leading to faster issue resolution and decreased equipment downtime. It also increased the accuracy of closure codes, improving maintenance logs and enabling better predictive model retraining. These improvements directly contribute to higher technician productivity and operational efficiency.

BEST PRACTICES

Start by identifying critical assets whose failure would disrupt operations. Install sensors to log vibration and temperature data and send it to the cloud. Gather OEM repair manuals for those assets and, if possible, collect historical maintenance records. This foundation ensures the AI assistant can deliver accurate, relevant, and context-aware diagnostic guidance.

OTHER APPLICATIONS

The solution is adaptable across industries where minimizing equipment downtime is essential, such as manufacturing, logistics, oil and gas, and healthcare. It can support predictive maintenance for motors, fans, gearboxes, bearings, conveyors, and other components, as long as sensors capture vibration and temperature data.