Quick Check

Intelligent Production Assistant on Edge (IPA Edge)

Ausgangssituation

Moderne Produktionsumgebungen in der diskreten Fertigung und in relevanten Bereichen der Prozessindustrie stehen unter hohem Effizienzdruck und leiden unter Fachkräftemangel. Mitarbeitende an den Maschinen müssen parallel die Produktion steuern, Parameter optimieren, Störungen beheben und Basiswartungen durchführen. Das dafür notwendige Wissen (Fehlercodes, Einstellwerte, Prozess- und Wartungsanleitungen) ist über umfangreiche Dokumentationen und verschiedene Systeme verteilt. Direkt an der Maschine ist der Zugriff häufig erschwert: Suchaufwände sind hoch, Netzwerkkonnektivität ist nicht immer gegeben, Datenschutzanforderungen limitieren die Cloud-Nutzung. Die Folgen sind Zeitverluste, vermeidbare Fehler, erhöhte Belastung und verlängerte Stillstandszeiten. Bestehende Wissensmanagement- und Supportprozesse bieten am Shopfloor nicht die nötige Geschwindigkeit und Verlässlichkeit. Vor diesem Hintergrund sollte im Quick Check die Grundlage für eine lokale, echtzeitfähige und datenschutzkonforme Unterstützung am Arbeitsplatz geprüft werden.

Lösungsidee

IPA Edge ist ein KI-Produktionshelfer, der lokal, cloud-unabhängig und ohne zusätzliche Inrastruktur auf einer SiMa-MLSoC-Plattform ausgeführt wird. Kern ist ein auf Produktionswissen spezialisiertes LLM (Dokumentationen, Prozessbeschreibungen, Fehlerkataloge, Wartungsanleitungen). Mitarbeitende interagieren per Sprach- oder Texteingabe; der Assistent erklärt Fehlercodes, gibt Schritt-für-Schritt-Anleitungen, empfiehlt Parameter und unterstützt bei Diagnosen und Basiswartung. Die Edge-Ausführung macht das System echtzeitfähig, offline-verfügbar und datenschutzkonform. Ziel: schnellere Problemlösung, reduzierte Stillstandszeiten, höhere Prozesssicherheit und Qualität sowie zentral zugängliches Know-how. Der Quick Check prüfte die LLM-Performance, 2-3 Kern-Use-Cases, Datenquellen, Integrationspfade und Akzeptanz.

In Zusammenarbeit mit

SiMa.ai

Umsetzung der KI-Applikation

Es wurden zwei mögliche Szenarien umgesetzt:

a) LLM-API auf dem SMA MLSC, Vektordatenbank extern (einfach, schnell)

- LLM läuft lokal als REST/gRPC-Service auf der Hardware.
- RAG über externe Vektordatenbank: Dokumente werden außerhalb gehostet, eingebettet und indexiert; Anfragen holen Top-k-Kontext aus der Datenbank und übergeben ihn ans lokale LLM.
- Vorteile: geringes Integrationsrisiko, schnelle Iteration, schnelles PoC. Nachteile: Netzabhängigkeit und potentielle Datenabflüsse je nach Hosting.

b) Gesamtsystem vollständig auf der Hardware mit embedded Vector Database

- LLM, Embedding-Modell und Vektordatenbank sind auf dem MLSC integriert.
- On-device Ingestion: Dateiimport, Chunking, Embedding und Indexaufbau; RAG vollständig offline.
- Optimierungen: Quantisierung, Streaming/Chunked Decoding, Speicher- und Thermomanagement, einfache Admin-UI/CLI für Konfiguration.
- Vorteile: Echtzeit, Offline-Fähigkeit, Datenhoheit/IP-Schutz; Nachteil: höherer Engineering- und Ressourcenaufwand.

»Gemeinsam mit dem Fraunhofer IPA konnten wir im Rahmen des Quick Checks einen Wartungs-ChatBot mit RAG auf unserer MLSoC-Plattform realisieren. Unsere im Markt für Physical AI führende Modalix SoC Plattform ermöglicht eine Steigerung der Produktivität in den Linien und bei der Einarbeitung von Mitarbeitern und hilft somit auch, den Fachkräftemangel zu adressieren.«

Stephan Reichenauer
SiMa.ai

Kontakt

Christof Nitsche
christof.nitsche@ipa.fraunhofer.de

Kontakt:
info@ki-fortschrittszentrum.de

Weitere Informationen:
www.ki-fortschrittszentrum.de

Fraunhofer-Institut für Produktionstechnik und Automatisierung IPA

Nobelstraße 12
70569 Stuttgart

www.ipa.fraunhofer.de