IT&Production E-PAPER
SONDERTEILE, BRANCHENSPECIALS, THEMENSCHWERPUNKTE



Bild: ©Aekkachai/stock.adobe.com

SONDERTEIL

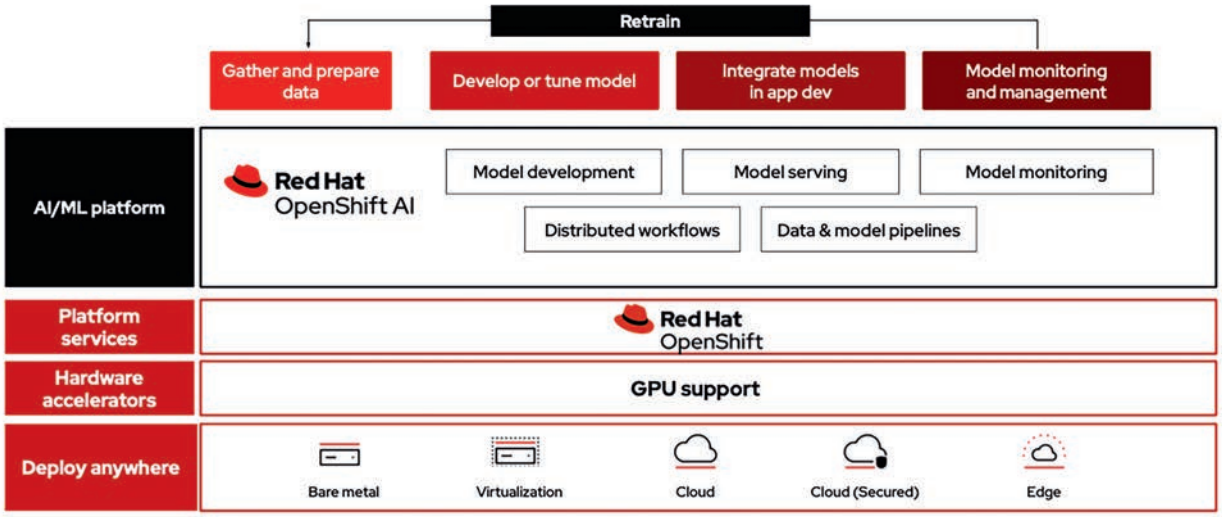
LARGE LANGUAGE MODELS

Support - Instandhaltung - Projekte

KI-gestützte Produktionsoptimierung

Generative KI ans Steuer lassen

Höre, was Du siehst!



Mit Red Hat OpenShift AI können Fertigungsunternehmen KI-gestützte Anwendungen in Hybrid-Cloud-Umgebungen entwickeln und bereitstellen.

Generative KI verspricht auch in der Produktion Vorteile. Unternehmen können damit die Effizienz und Qualität der Fertigung verbessern und dabei gleichzeitig die Maschinenbediener entlasten. Wenn man sie lässt.

KI ist in der Fertigungsindustrie kein neues Thema, wie die vielen Anwendungen von Predictive AI zeigen. Allerdings beschränkt sich der Einsatz von KI bisher oft nur auf die Maschinenebene. Der Zustand einer gesamten Produktionslinie oder die Kommunikation zwischen verschiedenen Systemen bleiben dabei häufig außen vor. Mit den Fortschritten der generativen KI wird hier aber ein Wandel einsetzen, der eine effizientere industrielle Fertigung maßgeblich vorantreiben wird. Das Potenzial der generativen KI zeigt sich gerade in der Mensch-Maschine-Interaktion. Dabei geht es nicht darum, den Maschinen-Operator zu ersetzen. Vielmehr soll er bei technischen Herausforderungen durch KI-generierte Hinweise unterstützt werden, um die Produktivität insgesamt zu erhöhen.

LLMs sind nicht alles

Wichtig bei der KI-Nutzung im Fertigungsbereich ist, nicht ausschließlich auf LLMs (Large Language Models) zu setzen. Diese Modelle sind mit sehr großen Datenmengen trainiert, das heißt, alle Informationen stecken in den Modellen und ihr Wissen ist immer die Basis für die Antworten. In der Regel sind somit keine aktuellen Daten vorhanden. Für die Echtzeitnutzung in der Produktion ist das folglich nicht ausreichend. Es geht also darum, dass nicht nur das LLM als Wissensquelle genutzt wird, sondern auch weitere Dokumente, Online-Web-Suchergebnisse und dann vor allem auch unternehmensspezifische Echtzeit- und Maschinendaten wie z.B. Logdaten oder Fehlermeldungen. Für die Datenspeicherung dienen dabei Vektordatenbanken. Aus den Daten generiert das LLM dann eine Antwort, die dem Maschinen-Operator hilft, Probleme zu lösen. Konkret wird dabei der Prompt des Users, das heißt des Maschinen-Operators, im Hintergrund mit den oben genannten Daten angereichert, sodass das LLM eine möglichst exakte Antwort gibt.

Schneller Forschritt bei Technik

Inzwischen gibt es bereits mehrere Ansätze und Techniken, die die LLMs mit weiteren Informationen anreichern. Bekannte Beispiele dafür sind RAG (Retrieval-Augmented Generation) und Agentic AI. RAG bietet die Möglichkeit, die in einem LLM vorhandenen Daten durch externe Wissensquellen zu ergänzen, etwa durch Daten-Repositories oder bestehende Dokumentation. So kann auch wesentlich einfacher sichergestellt werden, dass die Modelle richtige, sichere Antworten liefern, ohne zu halluzinieren. Mit AI Agents können Probleme unter Nutzung von Daten aus einem LLM,

Bild: Red Hat GmbH

einer Vektordatenbank, einer Knowledge Base oder aus dem Internet analysiert werden, um Schlussfolgerungen zu ziehen und Entscheidungen zu treffen, die dann als Informationen für einen Operator oder die eigenständige Veranlassung von Veränderungen dienen.

Kleine Modelle liegen im Trend

Bei der Auswahl des passenden Modells stehen Fertigungsunternehmen vor einer ersten Herausforderung. Hier zeichnet sich ab, dass immer mehr Firmen nicht auf LLMs, sondern auf kleinere Modelle setzen, die für spezielle Use Cases trainiert werden. Sie sind kosteneffizienter und mit geringerem Aufwand zu realisieren. Doch wie sollten Unternehmen nun eine Implementierung in Angriff nehmen – mit dem Ziel, KI nicht nur auf Maschinenbasis, sondern im Kontext einer gesamten Produktionslinie einzusetzen? Eine zentrale Voraussetzung ist die Verfügbarkeit der Daten. Im Großen und Ganzen sind sie vielfach schon vorhanden, etwa in Datenbanken oder Data Lakes. Eine Datenquelle bieten darüber hinaus die in der Fertigungsindustrie oft genutzten Digital Twins. Ist die Datenverfügbarkeit gegeben, geht es in einem nächsten Schritt um die semantische Verknüpfung der Daten und den Verbindungsaufbau zwischen den Maschinen. Hierfür eignet sich etwa eine Graphdatenbank.

Zwischen Cloud und Hausbetrieb

Wichtig für die KI-Nutzung ist zudem eine geeignete Infrastruktur, die die Einführung von der Entwicklung bis zum Produktivbetrieb vereinfachen und beschleunigen kann. Eine solche Infrastruktur stellt eine offene Container-basierte Hybrid-Cloud-Plattform bereit. Sie bietet zum einen eine Basis für die KI-Modell-Entwicklung, das KI-Modell-Training und auch die KI-Modell-Einbettung in Anwendungen. Zum anderen ist sie flexibel in einer Private-, Public- oder Edge-Umgebung nutzbar. Bei der KI-Nutzung stellt sich dann immer die Frage: Public Cloud oder On-Premises? In der Industrie gehen dabei derzeit viele Unternehmen den On-Premises-Weg. Dafür gibt es mehrere Gründe: Bei Public-Cloud-Services werden kontinuierlich neue Versionen ausgerollt, die beim industriellen Einsatz ein Problem darstellen können, da sich dann unter Umständen ein Gesamtsystem anders verhält. Bei der On-Premises-Nutzung eines Modells, das selbst versioniert und getestet wird, kann hingegen bei einer neuen Version relativ schnell und einfach der eigene Qualitätssicherungsprozess angewendet werden. Ein Self-Hosting unterstützt zudem bei der Erfüllung von Latenzanforderungen. Schließlich geht es gerade im Manufacturing um den Edge-Bereich, in dem KI-Algorithmen an der Datenquelle ausgeführt werden sollten.

Um Fertigungsprozesse effizient zu optimieren, müssen umfangreiche Datenmengen schnell analysiert werden, und zwar direkt an der Produktionslinie durch die Vernetzung der IT mit den Anlagen oder Steuersystemen. Aufgrund der Herausforderungen in puncto Antwortzeiten ist es hier nicht immer möglich, in die Public Cloud zu gehen.

Open Source Software

Nicht zuletzt sind Open-Source-LLMs derzeit ein viel diskutiertes Thema, da viele Unternehmen nachvollziehen wollen, was in den Parametern von LLMs enthalten ist und mit welchen Daten sie trainiert wurden. Mit Open Source ist das möglich. Dabei muss sich der Einsatz von KI in einem Industrieunternehmen keineswegs auf die reinen Fertigungsprozesse beschränken. Auch die Resilienz der Lieferketten kann beispielsweise verbessert werden, indem etwa durch Lieferantenanalysen und die Nutzung von AI Agents eine schnelle Umstellung von Lieferketten möglich wird. Das potenzielle Einsatzspektrum der generativen KI ist weitreichend – und deshalb wird sie ein entscheidender Treiber der weiteren industriellen Transformation sein. ■

Stefan Bergstein
Chief Architect Manufacturing
Red Hat GmbH
www.redhat.com

Anzeige



Die IT&Production berichtet über alle Aspekte der Industrie 4.0 und digitalen Transformation. Der E-Mail-Newsletter hilft Entscheidern die Herausforderungen der digitalen Produktion zu bewältigen. Jetzt Wissensvorsprung sichern:
www.it-production.com/newsletter-abonnieren

IT&Production
Zeitschrift für erfolgreiche Produktion

Wissensmanagement mit Agentic AI

KI-Agenten statt Tabellen scrollen

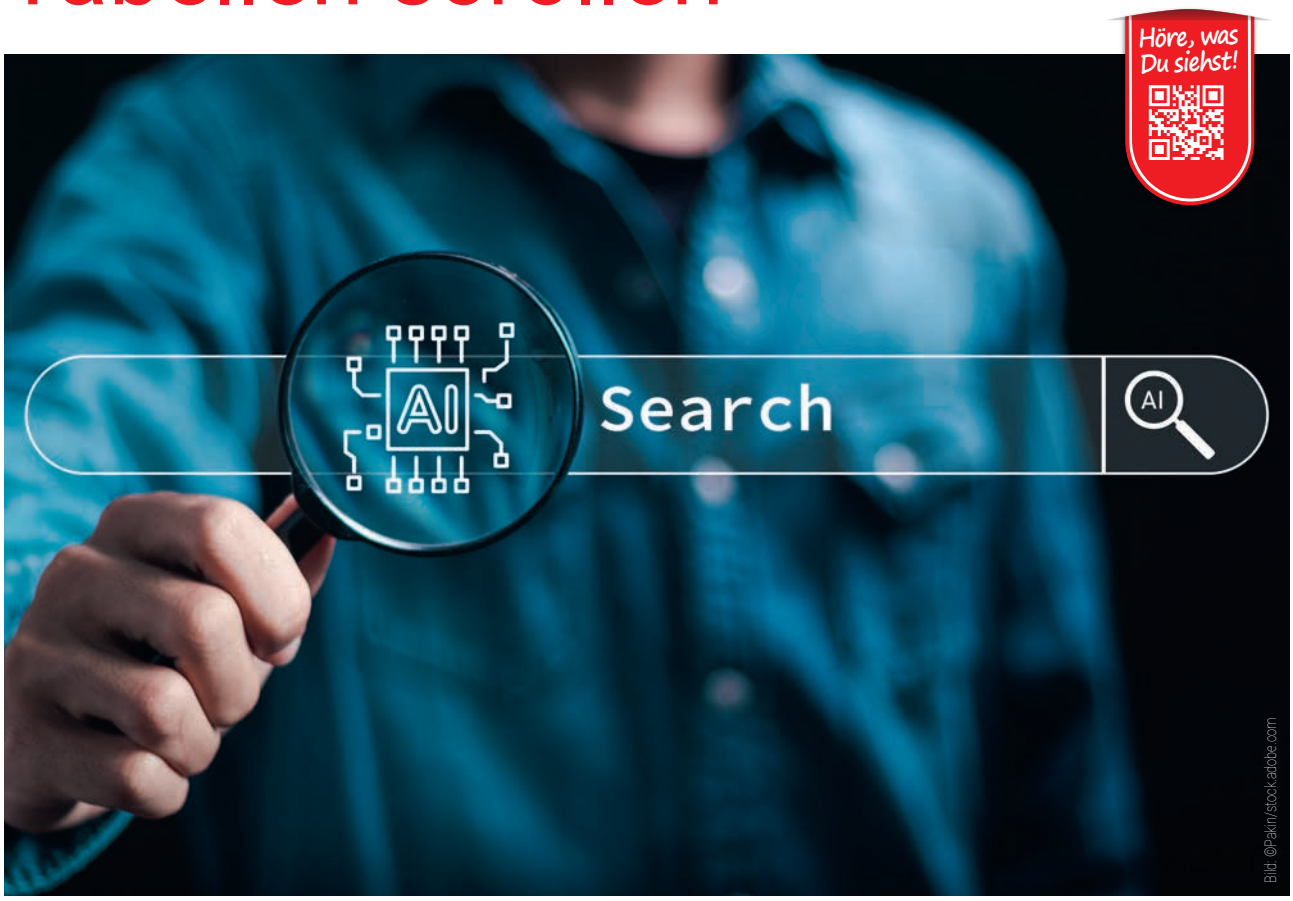


Bild: ©Pakin/stock.adobe.com

Viele produzierende Unternehmen setzen bereits auf KI-Anwendungen wie Predictive Maintenance oder generatives Design. Oft ein unerwünschter Nebeneffekt: Die vielen KI-Tools werden zu eigenen Datensilos. Hier können KI-Agenten glänzen, die Zugriff auf diverse Unternehmenssysteme haben und selbst komplexe Fragen der Mitarbeitenden in natürlicher Sprache beantworten.

Um das Potenzial von KI zu nutzen, braucht es intelligente Systeme, die Daten verknüpfen und effizient nutzbar machen. Hier kommen KI-Agenten ins Spiel, die entweder in den Anwendungen der Mitarbeitenden integriert sind, oder über den Browser aufgerufen werden können. KI-Agenten sind autonome Softwareprogramme, die in der Lage sind, selbstständig Aufgaben auszuführen und mit ihrer Umgebung zu interagieren. Sie kombinieren Methoden des maschinellen Lernens, der natürlichen Sprachverarbeitung und der Datenverarbeitung, um komplexe Anfragen zu beantworten und kontextbezogene Entscheidungen zu treffen. Laut der Boston Consulting Group können KI-Agenten hochkomplexe Problemlösungen übernehmen. KI-Agenten unterscheiden sich von herkömmlichen Chatbots dadurch, dass sie aktiv Informationen abrufen, verarbeiten und in geeigneter Form darstellen.

Einfach Fragen stellen

Die Handhabung ist einfach. Mitarbeitende stellen dem KI-Agenten Fragen wie ‘Wie ist der aktuelle Status von Projekt XY?’ – ganz wie in einer gewöhnlichen Mensch-zu-Mensch-Interaktion. Basierend auf den angebundenen Datenquellen sollte der Agent eine präzise Antwort liefern. Zusätzlich zeigt er weitere relevante Informationen an, wie etwa Risiken, Abhängigkeiten, Ressourcenauslas-

tung oder technische Dokumentationen. Der Agent wählt automatisch eine passende visuelle Darstellung. Abhängig vom Kontext sind die Informationen in Listen, Diagrammen oder Tabellen aufbereitet. Zudem gibt der KI-Agent Handlungsempfehlungen.

Wie KI-Agenten funktionieren

KI-Agenten basieren auf Large Language Models (LLMs), die ihre Hersteller mit großen generischen Textmengen trainiert haben. Diese Modelle verstehen die menschliche Sprache, generieren kontextbezogene Antworten und verarbeiten selbst komplexe Fragestellungen scheinbar mühelos. Doch um unternehmensspezifische und aktuelle Informationen einfließen zu lassen, reichen vortrainierte Modelle nicht aus. Hier unterstützen zusätzliche KI-Technologien:
Ein zentraler Ansatz beim Einsatz von KI-Agenten in Unternehmen ist **Retrieval-Augmented Generation (RAG)**. Diese Methode stellt sicher, dass die KI-Agenten nicht nur auf ihre Trainingsdaten zurückgreifen, sondern stets aktuelle und relevante Unternehmensdaten als Basis für die Beantwortung der Fragen heranziehen. Dadurch lassen sich Halluzinationen reduzieren.

Mindbreeze als Informationsdrehscheibe

Dabei dient die Wissensmanagement-Anwendung von Mindbreeze als Bindeglied zwischen dem LLM und den Unternehmensdaten. Mithilfe semantischer Suche, Natural Language Processing (NLP) und automatischer Klassifikation identifiziert das System die passenden Daten für das LLM. Sobald eine Anfrage eingeht, analysiert sie diese und durchsucht die angebundenen Datenquellen wie Wissensdatenbanken, Dokumentationen oder FAQs. Dabei ist es egal, ob die Daten strukturiert oder unstrukturiert vorliegen. Die Anwendung leitet relevante Informationen unter Berücksichtigung von Zugriffsrechten an das LLM weiter. Das LLM formuliert daraus eine Antwort und ergänzt sie mit Quellenangaben, sodass Mitarbeitende die Informationen bei Bedarf vertiefen können. Durch kontinuierliches Lernen aus den Unternehmensdaten und dem bisherigen Nutzerverhalten werden die Antworten mit der Zeit genauer und individueller.



Mario Matuschek ist AI Solution Manager bei Mindbreeze.

Bild: Mindbreeze GmbH

Konnektoren zur Datenanbindung

Damit ein KI-Agent effizient arbeitet, benötigt er Zugriff auf relevante Unternehmensdaten. Durch Konnektoren können Unternehmen ihre Daten unabhängig vom Speicherort – sei es auf lokalen Servern, in der Cloud oder auf Plattformen wie Microsoft Azure, Oracle, Google oder AWS – an Mindbreeze und damit an KI-Agenten anbinden. Die Daten bleiben dabei stets an ihrem ursprünglichen Speicherort. Das System greift in Echtzeit darauf zu. Dadurch bleibt die Datenverarbeitung konsistent, und Datenschutzrichtlinien werden eingehalten.

Deployment beeinflusst den Preis von KI-Agenten

Die Implementierung eines KI-Agenten erfolgt je nach IT-Infrastruktur eines Unternehmens unterschiedlich. Unternehmen mit hohen Datenschutzanforderungen und Kontrolle bevorzugen oft ein On-Premises-System. Alternativ ist der Einsatz als Software-as-a-Service (SaaS) möglich. Hier übernimmt der Anbieter Wartung und Updates. Eine weitere Option sind Marketplaces großer Cloud-Anbieter wie AWS, Azure oder Google Cloud. Diese ermöglichen eine schnelle Bereitstellung und Skalierung. Die Wahl der Deployment-Option hat direkten Einfluss auf die Kosten.

Praxisbeispiele für den Einsatz von KI-Agenten

Bei der **Produktentwicklung** sind zahlreiche Informationen aus Marktanalysen, früheren Projekten, technischen Spezifikationen und regulatorischen Vorgaben entscheidend. Ein KI-Agent durchsucht und strukturiert diese Daten automatisch, sodass Ingenieure und Designer schneller auf relevante Erkenntnisse zugreifen und fundierte Entscheidungen treffen.
Auch bei der **Optimierung der Produkte** leisten KI-Agenten Unterstützung. Sie analysieren Kundenfeedback, Felddaten und Wettbewerbsanalysen, um Verbesserungspotenziale zu identifizieren. So erkennen Mitarbeitende beispielsweise häufige Kundenbeschwerden frühzeitig, identifizieren schnell Fehlerquellen in der Technik und bewerten Designänderungen rascher hinsichtlich ihrer Auswirkungen auf Kosten und Effizienz. Die Einführung von KI-Agenten ist als wichtiger Schritt für Unternehmen zu verstehen, die wachsende Komplexität im Datenmanagement zu bewältigen. ■

Mario Matuschek
AI Solution Manager
Mindbreeze
www.mindbreeze.com

Scan mich!

Normungsroadmap künstliche Intelligenz

Wie weit ist die KI-Normung?



Die zweite Ausgabe der Deutschen Normungsroadmap KI hat den Bedarf an Normen und Standards für KI ermittelt und bildete den Auftakt für die Umsetzung.

Normen und Standards bieten Leitplanken beim Einsatz von künstlicher Intelligenz zur Optimierung der Produktion. Welche benötigt werden, hält die Normungsroadmap KI fest, die seit 2022 abgearbeitet wird. Doch wie viele Projekte sind in Arbeit? Welche sind abgeschlossen? Und was ist in Sachen KI-Normung noch zu tun?

Der Artificial Intelligence Act – kurz AI Act – ist die erste umfassende Verordnung der EU zur Regulierung künstlicher Intelligenz. Er soll sicherstellen, dass KI-Systeme innerhalb Europas verantwortungsvoll, sicher und im Einklang mit europäischen Werten eingesetzt werden. Die Verordnung zur KI-Regulierung weist Normen im Bereich der Hochrisiko-KI-Anwendungen eine zentrale Rolle zu. Anforderungen an KI-Systeme, u.a. an Transparenz, Robustheit und Genauigkeit, sollen durch harmonisierte Europäische Normen technisch konkretisiert werden. Eine zentrale Rolle spielt hierbei die Normungsroadmap KI: Sie formuliert Bedarfe für Normen und Standards, die in Umsetzung des AI Acts zu erarbeiten sind. Mehr als 570 Expertinnen und Experten waren bis Ende November 2024 an der Ausarbeitung des Gemeinschaftswerkes beteiligt. Das Bundesministerium für Wirtschaft und Klimaschutz (BMWK) hat das Projekt finanziert. Die Umsetzung der Empfehlungen aus der Roadmap ist nun in vollem Gange.

Neun Schwerpunkte

Die derzeit in Arbeit befindlichen oder bereits abgeschlossenen KI-Normungsprojekte sind neun Schwerpunktthemen zugeordnet, entsprechend der Unterteilung in der Normungsroadmap KI:

- Grundlagen
- Prüfung und Zertifizierung
- Sicherheit
- Soziotechnische Systeme
- Industrielle Automation
- Mobilität
- Medizin
- Finanzdienstleistungen
- Energie und Umwelt

Insbesondere in der industriellen Automation sind Normen und Standards essenziell: Sie fördern die unternehmensübergreifende Interoperabilität und helfen dabei, regulatorische Rahmenbedingungen umzusetzen. Beim Schwerpunktthema Industrielle Automation steht die digitale Transformation des produzierenden Gewerbes mittels KI im Fokus. Eine wichtige Funktion wird dabei der digitalen Abbildung der physischen Realität zugeschrieben, dem digitalen Zwilling. Die Normungsroadmap stellt Herausforderungen im Zusammenhang mit Datenmodellen für den Einsatz von KI in der industriellen Automation dar. Thema ist außerdem, wie Mensch und Maschine interagieren und wie sich KI-Systeme integrieren lassen.

Bilder der Doppelseite: DIN Deutsches Institut für Normung e.V.

Übergeordnete Empfehlungen

Zudem haben die Autorinnen und Autoren im Dokument sechs übergreifende Handlungsempfehlungen formuliert. Diese richten sich an Normung, Forschung und Gesetzgebung und verfolgen eine übergeordnete und langfristige Zielstellung:

- Horizontales Konformitätsbewertungs- und Zertifizierungsprogramm für vertrauenswürdige KI aufsetzen
- Dateninfrastrukturen und -qualitätsstandards zur Entwicklung und Validierung von KI-Systemen aufsetzen
- Den Menschen in allen Phasen des KI-Lebenszyklus als Teil des Systems begreifen
- Vorgaben für die Konformitätsbewertung lernender Systeme im Bereich der Medizin erarbeiten
- Über Best Practices sichere und vertrauenswürdige KI-Anwendungen in der Mobilität entwickeln
- Übergreifende Datenstandards und dynamische Modellierungsverfahren erarbeiten

Großteil erledigt oder in Arbeit

Welchen Status quo haben nun aktuell die Projekte? Von den 122 in der Normungsroadmap gelisteten Handlungsbedarfen richten sich 88 an die Normung, 14 an die Forschung und neun an die Gesetzgebung. Die Umsetzung der 88 Bedarfe im Bereich Normung ist auf einem guten Weg: Ein Großteil der zugehörigen Normungsprojekte ist erledigt oder in den entsprechenden Gremien in Arbeit. So wurden Stand Dezember 2024 bereits 54 Prozent umgesetzt bzw. teilweise umgesetzt. Weitere 29 sind derzeit in Umsetzung, elf Prozent in Vorbereitung. Ein guter Zwischenstand, der veranschaulicht, wie effizient und zügig sich wichtige Normen und Standards für eine sichere und vertrauenswürdige KI 'Made in Germany' realisieren lassen. Ein Nebeneffekt in diesem Zusammenhang: Bei der Umsetzung der Normungsroadmap KI wurden – Stand Dezember 2024 – 346 KI-Fachleute an die Normung herangeführt, davon 85 für die Mitarbeit in Normungsgremien und 261 für die Mitarbeit in Arbeitsgruppen wie DIN SPEC-Gremien, Ad hoc-Gruppen und Anwenderkreisen. Engagieren sich diese Expertinnen und Experten aus Wirtschaft, Wissenschaft, Politik und Zivilgesellschaft auch künftig in der KI-Normungsarbeit, ist das eine nachhaltige Investition in eine sichere und wertebasierte KI aus Deutschland und Europa. Denn zügig im Konsens erarbeitete, markt- und bedarfsgerechte Normen und Standards sind Basis für innovative Systeme, die diese Technologie nutzen.



Umsetzungsstand der Normungsbedarfe aus der zweiten Ausgabe der Deutschen Normungsroadmap KI (Stand: Dezember 2024, www.din.de/go/umsetzung-roadmap-ki).

KI im Engineering

Folgende Beispiele aus dem Bereich der industriellen Automation illustrieren, welche Themen und Normungsbedarfe aus der Normungsroadmap KI umgesetzt wurden. So hatten die Autoren und Autorinnen einen Bedarf für ein Referenzmodell für KI-Engineering ermittelt. Ziel war es, Begriffe und Konzepte sowie deren Zusammenhänge zum System Engineering unter besonderer Berücksichtigung des Einsatzes von KI-Methoden in Subsystemen zu definieren und zu erläutern. Und so für Ingenieurinnen, Informatiker und Datenwissenschaftlerinnen ein gemeinsames Grundverständnis zu schaffen. Dieser Bedarf wurde über mehrere Normen umgesetzt, etwa über die ISO/IEC20546 'Informationstechnik – Big Data – Überblick und Begriffe' und die ISO/IEC22989 'Informationstechnik – künstliche Intelligenz – Konzepte und Terminologie der künstlichen Intelligenz'.

Von Modellen und Metriken

Ein weiteres Beispiel: Ermittelt wurde ein Bedarf an validierbaren Transformationen von Strukturen und Modellen. Transformationsmechanismen von Werkzeugen und Systemen für den Import und Export von Strukturen und Modellen sollten transparent und überprüfbar sein, um Veränderungen der transformierten Inhalte erkennen zu können sowie Fehlinterpretationen zu vermeiden. Dies wurde über die Normungsprojekte DIN EN ISO/IEC12792 und ISO/IEC TS 8200 realisiert. Ebenso gefragt war eine Definition von Metriken und Methoden, um die Datenqualität u.a. in Machine-Learning-Datenmodellen bewerten zu können. Denn die Datenqualität ist ein entscheidender Faktor, sobald Transaktionen über Datenmodelle ausgeführt werden. Umgesetzt wurde dies über die Norm ISO/IEC5259 'Künstliche Intelligenz – Datenqualität für Analytik und maschinelles Lernen (ML) – Teil 2: Datenqualitätsmaßnahmen'. Weitere KI-Normungsprojekte im Bereich Industrielle Automation sind noch in Umsetzung oder wurden teilweise umgesetzt. Zwei Projekte sind noch unbehandelt, eines ist in Vorbereitung: der Aufbau einer standardisierten Metadatenbeschreibung von KI-Methoden. ■

Tipp:

Einen aktuellen Überblick über die erledigten und noch in Arbeit befindlichen Themen gibt es unter *www.din.de/go/normungsroadmapki*

Filiz Elmas-Arslan
Leiterin Strategische Entwicklung Künstliche Intelligenz
DIN Deutsches Institut für Normung e.V.
www.din.de

Leitplanken für große Sprachmodelle

Wie Retrieval Augmented Generation LLMs unterstützt



Bild: ©Antony Weerut/stock.adobe.com

Ob Bedienungsanleitungen für Autos oder Rechtstexte – entscheidende Stellen zu finden, kann dauern. Und ob man am Ende auch alle relevanten Einträge gefunden hat? Ein Team am Fraunhofer IWU hilft mit einer KI auf die Sprünge, 'füttert' sie mit umfangreichen Fachtexten und bereitet den externen Input so auf, dass Suchanfragen (Prompts) zu präzisen und erschöpfenden Informationen führen sollen. Retrieval Augmented Generation (RAG) macht es möglich.

Large Language Models (LLM) bzw. große Sprachmodelle liefern in Chatbots auf alltägliche Anfragen häufig gute bis sehr gute Ergebnisse, die oft für eine erste Orientierung mehr als ausreichend sind. Forschende am Fraunhofer IWU geben allerdings zu bedenken, dass Trainingsdatensätze unvollständig oder veraltet sowie einige Informationen unscharf oder gar falsch sein können. Sie raten daher zu einer Überprüfung der erhaltenen Auskünfte. So sollte man sich etwa bei Rechtsfragen nicht 'blind' auf einen Chatbot verlassen. Und was, wenn von den Anga-

ben die Sicherheit von Menschen abhängen kann? Doch wieder selbst eingehend relevante Dokumente studieren?

Retrieval Augmented Generation

Mit Retrieval Augmented Generation (RAG) wollen Forschende am Fraunhofer IWU hier Abhilfe schaffen und zusätzliche Leitplanken für das Sprachmodell schaffen. Das LLM durchleuchtet dann in erster Linie maßgebliche Texte bzw. Textstellen. Das Sprachmodell wird dabei nicht neu trainiert, sondern selektiv erweitert.

Betrieb auch auf Standard-PCs

Als Sprachmodell verwenden die Forschenden LLaMA (Large Language Model Meta AI). Dieses sei groß bzw. leistungsfähig genug und überfordere dennoch nicht Rechenleistung und Grafikkarte eines hochwertigen Standard-PCs. Die Anwendung kann somit auf einem lokalen Rechner betrieben werden und Unternehmen behalten die Hoheit über ihre Daten. Bei weniger sensiblen Daten ist auch der Betrieb in der Cloud möglich.

Wie funktioniert RAG?

Zunächst müssen die ins LLM zu importierenden Daten auf den reinen Text reduziert werden (Cleaning). Sobald dieser in kleinere Abschnitte (Chunks; auffindbare Bausteine) segmentiert ist, folgt der Aufbau eines Suchsystems (Retrieval System), das die Chunks effizient durchsuchen kann. Die Chunks werden nach relevanten Passagen gegliedert und in einer Vektordatenbank abgelegt, also in mathematische Vektoren umgewandelt, die ihre Bedeutung repräsentieren. Auch die Prompts werden in Vektoren umgewandelt. So wird das Modell in die Lage versetzt, nach den in der Anfrage enthaltenen Wörtern zu suchen und gleichzeitig den Prompt tatsächlich zu 'verstehen' (semantische Suche). Das Modell kann jetzt die zu einer Nutzeranfrage passenden Chunks kürzen, neu strukturieren, die wichtigsten Informationen herausfiltern und zu einem verständlichen Zusammenhang kombinieren. Liegt eine konkrete Suchanfrage vor, stehen ausgewählte Chunks zur Verfügung, auf deren Grundlage das Modell faktenbasierte Antworten geben kann. Das Modell nutzt den zusätzlichen Kontext der Chunks und muss nicht neu trainiert werden.



Die Qualität der Antworten von Chatbots, die auf großen Sprachmodellen basieren, hängt von den Trainingsdaten ab: Grundlage sind meist eine Fülle öffentlich zugänglicher Dokumente. Die Verlässlichkeit der Antworten sollten Nutzerinnen und Nutzer kritisch hinterfragen.

Die richtige Strukturierung

Werkzeugmaschinen zählen zu den Kernkompetenzen des IWU. Das Team 'Maschinelles Lernen in der Produktion' weiß daher, auf welche Filter und Vorstrukturierungen es ankommt, damit die jeweils entscheidenden Passagen der Maschinenverordnung erreicht werden. Ein wichtiges Augenmerk legen die Forschenden künftig außerdem auf weitere Datenquellen wie etwa Tabellen oder Bilder, die so integriert werden müssen, dass sie gut auffindbar sind.

EU-Maschinenverordnung als Test

Die Maschinenverordnung soll für einheitliche Standards innerhalb der EU sorgen. Sie definiert etwa, wann bei Änderungen an Maschinen und Anlagen eine neue Konformitätsbewertung erforderlich ist. Als Beispiel für einen anspruchsvollen Rechtstext macht sie den Anfang bei den Demonstrationsanwendungen am Institut. Das Fraunhofer-Team will interessierten Firmen helfen, weitere Anwendungen aufzubauen. Infrage kommen dafür etwa Bedienungsanleitungen, die Angebotserstellung oder die Automatisierung von Programmierungsarbeiten im Produktionskontext ebenso wie Berichtspflichten. Das Durchsuchen betriebsinterner Dokumente, Regelungen, Betriebsvereinbarungen oder operativer Daten kann künftig ebenfalls KI-gestützt erfolgen. ■

Fraunhofer-Institut für Werkzeugmaschinen und Umformtechnik IWU
www.iwu.fraunhofer.de

Scan mich!

Vorlagen für Brainstormings

KI-Anwendungsfälle IDENTIFIZIEREN

Für KI-Tools gibt es viele Anwendungen. Doch wie findet man die nützlichsten? Jens-Uwe Meyer, Vorstandsvorsitzender von Innolytics, hat zehn KI-Templates entwickelt, die beim Brainstorming etwa in Meetings oder mit ChatGPT die Gedanken in eine zielführende Richtung lenken.



Jedes Template besteht aus einer Frage, die jeweils bezogen auf eine Organisation, ihre Zielgruppe, ihren Anwendungsfall oder Ziele spezifiziert werden müssen.

1. Fehlertemplate

Fehlererkennung
Durch KI können wir Fehler *(wo genau?)* erkennen und dadurch *(was erreichen?)*

Künstliche Intelligenz kann wie ein Detektiv Fehler aufspüren – etwa indem sie mit optischen Sensoren Prozesse und Datenströme überwacht. Bei einer Soll-Ist-Abweichung schlägt das KI-System Alarm. So hilft es Unternehmen, Schwachstellen etwa in der Fertigung, Lagerhaltung oder Kundenbetreuung zu erkennen und zu beheben.

2. Prognosetemplate

Prognosen
Indem KI Prognosen über *(was genau?)* erstellt, können wir *(was erreichen?)*

Künstliche Intelligenz kann auf der Basis von Musteranalysen Vorhersagen treffen. Sie kann prognostizieren, wie stark Zielkunden unter Voraussetzungen wie Wetter, Fördermittel usw. gewisse Leistungen nachfragen oder wie wahrscheinlich eine Maschine unter bestimmten Parametern bald Probleme bereitet. Das Prognosetemplate hilft bei der Bewertung, was diese Vorhersage überhaupt nutzt.

3. Beschleunigungstemplate

Beschleunigung
Wir sind effizienter, wenn wir *(welche Aufgabe?)* im *(wo genau?)* automatisieren

Das Beschleunigungstemplate hilft, Anwendungsfälle für Prozessautomatisierungen zu identifizieren. Dafür werden die Tätigkeiten eingetragen, die im Betrieb noch manuell ausgeführt werden. Wenn der Prozess in möglichst viele Teilaufgaben zerlegt wird, werden die Ergebnisse besser.

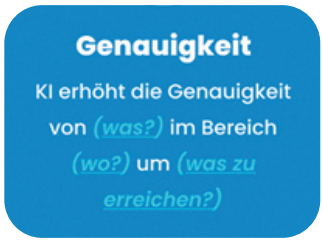
4. Individualisierungstemplate

Individualisierung
Wenn wir *(wen genau?)* gezielter ansprechen, können wir *(was erreichen?)*

Viele Menschen werden heute von Informationen überflutet. Deshalb scannen sie diese in der Regel nur kurz und entscheiden dann, ob und wie sie darauf reagieren. Dabei gilt für fast alle Kommunikationsprozesse: Individuelle Botschaften erreichen ihre Empfänger besser. Doch wo lohnt sich der anfallende Mehraufwand dafür? Diese Frage, das zu beantworten.

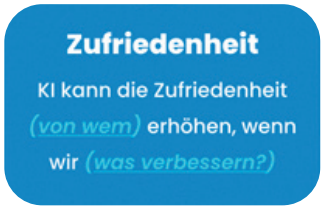
Bild ©Murrstock/stock.adobe.com

5. Genauigkeitstemplate



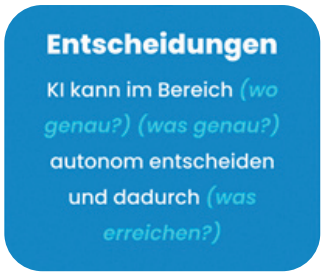
Entscheidungen fallen oft intuitiv aufgrund von Wissen und Erfahrung. Entsprechend risikobehaftet sind sie dann auch. Gibt es Bereiche, in denen eine Analyse und Erhöhung der Genauigkeit bessere Ergebnisse ergibt? Zuweilen wäre ein KI-Tool hilfreich, das beispielsweise technische Daten oder Marktdaten erfasst und auswertet. Doch wie groß wäre der betriebswirtschaftliche Nutzen des Tools? Das Genauigkeitstemplate hilft bei der Einschätzung.

6. Zufriedenheitstemplate



In der Zusammenarbeit mit Kunden, Partnern, Mitarbeitern und Lieferanten können Schwierigkeiten auftreten, wenn Infos nicht geliefert werden können, sich Termine verzögern oder die Leistungen nicht passen. Die Zufriedenheit sinkt. Das Zufriedenheitstemplate hilft, Anwendungsfälle zu identifizieren, die dazu beitragen, die Zufriedenheit der verschiedenen Interessengruppen zu erhöhen.

7. Entscheidungstemplate



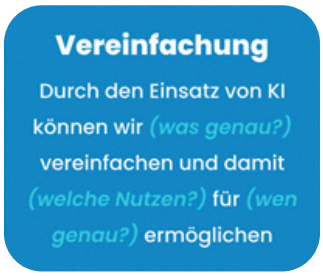
In Unternehmen sind Entscheidungsprozesse oft langwierig. Je mehr Personen und Abteilungen involviert sind, umso komplexer und somit aufwendiger und teurer wird dieser in der Regel. Mit dem Entscheidungstemplate lassen sich Anwendungsfälle im Betrieb identifizieren, wo sich Entscheidungen durch Algorithmen und KI-Systeme wirtschaftlich automatisieren lassen.

8. Rentabilitätstemplate



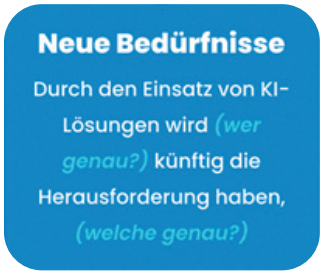
Viele Leistungen sind für Zielkunden durchaus interessant, ihre Erbringung für Unternehmen aber etwa aufgrund ihres Personalaufwands oder nötiger technischer Infrastruktur nicht. Mit KI-Einsatz könnte sich das ändern. Das Rentabilitätstemplate hilft bei der Suche nach Leistungen und Geschäftsmodellen, die sich bislang nicht rechneten, durch KI-Einsatz aber lukrativ werden könnten.

9. Vereinfachungstemplate



Während das Freistellen von Objekten in Photoshop traditionell anspruchsvoll ist, genügt auf dem iPhone, das Objekt mit dem eine halbe Sekunde lang zu markieren – schon ist es ausgeschnitten. KI kann Anwendungen deutlich vereinfachen. Mit dem Vereinfachungstemplate lassen sich Anwendungsfälle ermitteln, in denen Vereinfachungen neue Kunden gewinnen oder die Zufriedenheit bestehender Kunden steigern.

10. Bedürfnistemplate



Die Verbreitung von KI-Anwendungen in der eigenen, aber auch anderen Organisationen verändert Bedürfnisse und Erwartungen, auch bei Kunden und Interessanten. Veränderungen bergen immer Risiken, aber auch Chancen. Also sollten Unternehmer überlegen, bei welchen Kunden durch die KI neue Bedürfnisse und somit Chancen entstehen. Dabei hilft das Bedürfnistemplate.

Anwendung der Templates

Die Templates dienen dem Prozess der Ideengenerierung für einen potenziellen KI-Einsatz im Betrieb. Dieser sollte dadurch systematischer und strukturierter ablaufen. Mit steigender Qualität, weil bestimmte Bereiche und Anwendungsfälle fokussiert werden. So mancher Vorschlag, den etwa ChatGPT unterbreitet, wird banal und oberflächlich sein. Manager und Unternehmer müssen die Ergebnisse also noch immer bewerten, modifizieren und als Grundlage für eigene Überlegungen nutzen. Wunder vollbringt generative KI nicht. ■

Hinweis: Die im Artikel beschriebenen Templates sind auch in dieser kostenlosen Veröffentlichung zur KI-Potenzialanalyse enthalten: www.ki-roadmap.de

Jens-Uwe Meyer
CEO
Innolytics AG
www.innolytics.de