Fraunhofer
BIG DATA AI



Jenseits von Text und Bild – Generative KI für vielfältige Datenwelten

Praxisnahe Einblicke und innovative Anwendungsbeispiele zur Erschließung komplexer Unternehmensdaten

Fraunhofer Verlag

Fraunhofer-Allianz Big Data und Künstliche Intelligenz

Jenseits von Text und Bild – Generative KI für vielfältige Datenwelten

Praxisnahe Einblicke und innovative Anwendungsbeispiele zur Erschließung komplexer Unternehmensdaten

Fraunhofer Verlag

Vorwort

Generative Künstliche Intelligenz hat sich in kürzester Zeit zu einer der bedeutendsten zukunftsweisenden Technologien entwickelt. Sie wird die Wirtschaft in den kommenden Jahren nachhaltig transformieren und unsere Arbeitsprozesse in fast allen Branchen grundlegend verändern.

Generative KI (engl.: Generative Artificial Intelligence, kurz: GenAI) ist in der Lage, komplexe Probleme zu analysieren, inhaltlich zu interpretieren und Lösungsvorschläge zu generieren. In Zeiten von Fachkräftemangel und Wettbewerbsdruck kann GenAI für Unternehmen ein mächtiges Werkzeug sein: insbesondere in Form von KI-Anwendungen, die in den Bereichen Prozessautomatisierung, Produktivitätssteigerung, Innovationsmanagement oder bei der Einhaltung regulatorischer Vorgaben unterstützen.

Bekannt und beliebt geworden sind GenAI-Modelle, weil sie beeindruckende Ergebnisse in der Text- und Bildgenerierung erzielen. Doch in Unternehmen liegen Daten häufig nicht in Form von Bildern oder Texten vor, sondern haben spezielle Formate wie etwa Tabellen, Messdatenreihen, Logdateien oder Geodaten. Möglicherweise handelt es sich sogar um Prozess-, Bewegungs- oder Steuerungsdaten von Robotern oder um Pläne und Strukturbeschreibungen aus Digitalen Zwillingen. Die Verarbeitung und Nutzung dieser komplexen und vielfältigen Daten, die wertvolle Informationen und Prozesse enthalten und die sehr häufig den zentralen Kern der Wertschöpfung des Unternehmens ausmachen, stellt für viele Unternehmen eine große Herausforderung dar.

Für die Verarbeitung komplexer und anwendungsnaher Unternehmensdaten jenseits von Bildern und Texten entstehen aktuell zahlreiche spezialisierte, oder multimodale GenAI-Modelle, die ein ganz neues Spektrum an Möglichkeiten eröffnen.

Das vorliegende Whitepaper bietet einen Überblick über neue Ansätze zur Erschließung solch komplexer Unternehmensdaten mithilfe von generativer KI. Unsere Autorinnen und Autoren aus den Instituten der Fraunhofer-Allianz Big Data und Künstliche Intelligenz zeigen anhand konkreter Anwendungsbeispiele, wie diese Daten kontextuell interpretiert und zur Optimierung und Automatisierung von Prozessen genutzt werden können. Zahlreiche praktische Beispiele veranschaulichen das Potenzial, die Einsatzmöglichkeiten und Leistungsfähigkeit von GenAI in verschiedenen Unternehmenskontexten und Datenwelten jenseits von Text und Bild.

Lassen Sie sich von uns inspirieren: Entdecken Sie in diesem Whitepaper neue Ansätze, um Ihre Datenanalysen zu optimieren. Gehen Sie die nächsten Schritte, um GenAI in Ihrem Unternehmen erfolgreich anzuwenden und Ihre Innovationspotenziale zu heben.

Mit herzlichen Grüßen

Dr. Sonja Holl-Supra und **Prof. Dr. Dirk Hecker**
Geschäftsführung der Fraunhofer-Allianz Big Data und Künstliche Intelligenz

Weitere Informationen und vertiefende Praxisbeispiele zu den Anwendungsfällen finden Sie auf unserer Webseite:
www.bigdata-ai.fraunhofer.de/whitepaper_genai

Inhalt

Einleitung **5**

1. Tabellarische Daten und Matrizen **12**
1.1. Die Datenstruktur 12
1.2. Warum GenAI für diese Datenstruktur? 12
1.3. Use Cases 15
1.3.1. Voraussage von strukturierten Geschäftsdaten 15
1.3.2. Synthetisierung von Quantenschaltungen 16

2. Zeitreihen **17**
2.1. Die Datenstruktur 17
2.2. Warum GenAI für diese Datenstruktur? 17
2.3. Use Cases 20
2.3.1. Assistenzsystem zur Parameterwahl für Produktionsprozesse 20
2.3.2. Eventdaten als Treiber für moderne Process Intelligence 21
2.3.3. Large-Timeseries-Modelle zur Prognose von Energiezeitreihen 22
2.3.4. Agentengestützte interaktive Datenanalyse 23
2.3.5. Pretrained Transformer-Modelle zur funkbasierten Lokalisierung 24
2.3.6. Radarbasierte Objekterkennung und -klassifikation 25
2.3.7. Sparsame Zustandsüberwachungen in industriellen Maschinenparks 26

3. Graphen für komplexe Netzwerke **27**
3.1. Die Datenstruktur 27
3.2. Warum GenAI für diese Datenstruktur? 27
3.3. Use Cases 30
3.3.1. Sprachmodelle für die Anlagensteuerung und -überwachung 30
3.3.2. Lernassistent für die Personalisierung in Lernplattformen 31
3.3.3. De-Novo-Molekulardesign für gezielt ausgelegte Substanzen 32
3.3.4. Prognose von Verkehrsmengen für die Standortplanung 33
3.3.5. Eine flexible Datenstruktur für effiziente Schiffsüberwachung 34
3.3.6. Flexibles Management für Telekommunikationsnetzwerke 35

4. Digitale Zwillinge **36**
4.1. Die Datenstruktur 36
4.2. Warum GenAI für diese Datenstruktur? 37
4.3. Use Cases 39
4.3.1. Digitale Zwillinge für effiziente Gebäudebewirtschaftung 39
4.3.2. Digitaler Zwilling von Fahrzeuginsassen 40
4.3.3. Der digitale Produktpass 41
4.3.4. Semantisches Datenmapping bei Digitalen Zwillingen 42

5. Multimodale Daten und Robotik 43
5.1. Die Datenstruktur 43
5.2. Warum GenAI für diese Datenstruktur? 43
5.3. Use Cases 45
5.3.1. Steuerung von autonomen Baumaschinen 45
5.3.2. Robotik und Anlagenprozesse mit Sprach- und KI-Agenten steuern 46
5.3.3. Kollaborative Robotik mit Pose- und Gaze-Tracking 47
5.3.4. Revolutionierung der Kommissionierung: GenAI für die Logistik 48
5.3.5. Intelligente Produktbeschreibung bei Retouren und Re-Commerce 49

6. Fazit und Ausblick 50
6.1. Management Summary 50
6.2. Die Zukunft in Unternehmen mit GenAI 51
6.3. GenAI als Gamechanger – Empfehlungen für die Erschließung komplexer Unternehmensdaten 52

Referenzen 54

Autorinnen und Autoren 57

Abbildungsverzeichnis 59

Einleitung

Mit GenAI Herausforderungen begegnen und Chancen entdecken

Unternehmen stehen täglich vor vielfältigen Herausforderungen, um ihre Wirtschaftlichkeit und Wettbewerbsfähigkeit in einer zunehmend digitalisierten und dynamischen Marktlandschaft zu sichern. Die Fähigkeiten, komplexe Unternehmensdaten zu verstehen, verwertbare Einsichten zu gewinnen, datengestützte Entscheidungen zu treffen oder Automatisierungen zu implementieren, sind entscheidende Faktoren für den Erfolg.

Seit den frühen 2010er-Jahren haben technologische Fortschritte wie leistungsfähigere Hardware, große Datenmengen und verbesserte Algorithmen die Voraussetzungen in Unternehmen für datengetriebene Lösungen geschaffen. Wurden davor klassische Methoden des Maschinellen Lernens auf strukturierten Daten genutzt, um Muster zu erkennen, Vorhersagen und datengestützte Entscheidungen zu treffen oder Prozesse zu automatisieren, so ermöglichten nunmehr sogenannte Deep-Learning-Verfahren mit tiefen neuronalen Netzen die Verarbeitung unstrukturierter Daten. Die 2017 veröffentlichten Transformer-Modelle gaben dann den Anstoß zur generativen KI (GenAI), der Ende 2022 mit ChatGPT der Sprung aus der Forschungs- und Experimentierphase in die breite Masse gelang. Seitdem etabliert sich GenAI als Schlüsseltechnologie, die Unternehmen in allen Branchen Vorteile verschafft (Potenziale Generativer KI für den Mittelstand) [1].

GenAI kann, wie die vorausgegangenen Maschinellen Lernverfahren Daten analysieren und Vorhersagen treffen. Darüber hinaus – und darin liegt ihre besondere Innovationskraft – kann sie umfangreiche Inhalte (sogenannte Kontexte) erfassen und eigenständig neue, qualitativ hochwertige und kreative Inhalte und Lösungsvorschläge erzeugen. Sie kann außerdem komplexe, mehrschrittige Aufgaben lösen und vielfältige Arbeitsabläufe automatisieren. Damit öffnet GenAI für uns Menschen den Zugang zu ganz neuen Möglichkeiten, wie wir Daten analysieren und auf dieser Basis wichtige Entscheidungen treffen können. Ganz wichtig: Durch die Verwendung von vortrainierten GenAI-Modellen werden deutlich weniger unternehmensspezifische Daten benötigt als bei früheren Ansätzen. Dies ermöglicht eine sehr viel schnellere und in manchen Fällen erstmalige Umsetzung von Anwendungsideen.

GenAI kann komplexe Datenwelten erschließen

In den Prozessen von Unternehmen und Organisationen geht es allerdings nicht allein um die Verarbeitung von Bildern und Texten. Die Daten können sehr komplex und heterogen sein: Ob Messdatenreihen in der Produktion, Molekülstrukturdaten in Pharmazie oder Chemie, Tabellen im Finanzsektor, Anlagenstrukturdaten in Digitalen Zwillingen oder multimodale Sensordaten in Robotern – Daten liegen oft in besonderen Formaten vor oder benötigen eine spezielle Handhabung.

Tabelle 1: Die Möglichkeiten von GenAI in Unternehmen aus verschiedenen Branchen

Nutzen	Beschreibung	Beispiele/Anwendung
Automatisierung & Effizienzsteigerung	Unternehmen setzen GenAI ein, um repetitive Aufgaben zu automatisieren, was zu erheblichen Zeit- und Kosteneinsparungen führt.	Automatisierte Texterstellung, Bildgenerierung, Chatbots im Kundenservice sowie Prozessautomatisierung in der Fertigung.
Produktivitätssteigerung	GenAI-Tools unterstützen Mitarbeitende, um schneller und zielgerichteter zu arbeiten.	Anwendungen wie ChatGPT, Copilot und unternehmensspezifische Lösungen ermöglichen automatische Zusammenfassungen und die Beantwortung von Support-Anfragen.
Innovation & neue Geschäftsmodelle	GenAI beschleunigt die Entwicklung neuer Produkte und ermöglicht die Schaffung personalisierter Angebote für Kundinnen und Kunden.	Beispiele hierfür sind maßgeschneiderte Produktempfehlungen im Einzelhandel, individuelle Behandlungspläne im Gesundheitswesen und innovative Service-Modelle.
Qualitätsverbesserung	Durch den Einsatz automatisierter Qualitätskontrollen können Unternehmen die Fehlerquote reduzieren und kreative, datenbasierte Marketingkampagnen entwickeln.	Anwendungen umfassen die Qualitätskontrolle in der Fertigung, KI-gestützte Marketingkampagnen sowie die Analyse von Kundenbewertungen zur Verbesserung von Produkten und Dienstleistungen.
Forschung & Entwicklung	GenAI ermöglicht eine schnellere Entwicklung neuer Medikamente und unterstützt die effiziente Analyse großer Datenmengen zur Generierung neuer Hypothesen.	Beispiele sind die Medikamentenentwicklung im Gesundheitswesen, umfassende Datenanalysen in der Forschung und die Durchführung von Simulationen und Prognosen.

Angepasste oder spezialisierte GenAI-Modelle bieten vielfältige Möglichkeiten und großes Potenzial für das Kerngeschäft von Unternehmen, etwa in der Produktion von Automobilteilen oder der Forschung und Entwicklung von Medikamenten (siehe Tabelle 1). Anwendungen dieser Modelle können tief in Unternehmensprozesse integriert werden, um präzisere und mehrschrittige Antworten zu liefern, spezifische Muster und Trends zu identifizieren und schnellere Antwortzeiten zu ermöglichen. Dadurch tragen sie wesentlich zur Steigerung der Effizienz und Förderung der Innovationskraft bei.

Im vergangenen Jahr wurden spezialisierte Modelle maßgeblich durch wissenschaftliche Aktivitäten vorangetrieben und in der Regel als Open Source zur Verfügung gestellt. Solche Modelle sind nicht nur transparenter (d. h. sie bieten Informationen über Training und Daten, sind nachvollziehbar und erklärbar), sondern lassen sich auch umfassender anpassen. Dies kommt wiederum dem Schutz sensibler Daten zugute und stärkt die technologische Souveränität Deutschlands und Europas. Durch den On-Premise-Betrieb oder zumindest einen Betrieb in Deutschland haben Unternehmen die Kontrolle über Modell und Daten und werden unabhängig von unkalkulierbaren Änderungen der Technologieanbieter und -betreiber.

Aufbau und Gliederung des Whitepapers
Mit praxisnahen Anwendungsbeispielen zeigt dieses Whitepaper auf, wie Unternehmen ihre Effizienz und Wettbewerbsfähigkeit entscheidend steigern können, indem sie mit speziell zugeschnittenen GenAI-Modellen komplexe Daten effektiv erschließen und Prozesse optimieren oder automatisieren.

Eine kompakte Einführung gibt zunächst einen Überblick über die wichtigsten Konzepte von GenAI und die wichtigsten technischen Hebel.

Die fünf anschließenden Hauptkapitel widmen sich jeweils einem spezifischen Datentyp. Sie stellen zudem in Form von Anwendungsbeispielen aus verschiedenen Branchen technische Ansatzmöglichkeiten, Voraussetzungen und Vorteile vor.

Das Whitepaper stellt abschließend in einem Ausblick zukünftige Entwicklungen sowie Möglichkeiten dar, und zeigt, wie Unternehmen von den Fortschritten von GenAI auch in der Zukunft profitieren können.

Das erste Kapitel befasst sich mit **tabellarischen Daten** und beschreibt, wie tabellenspezifische Transformer-Modelle fundierte Prognosen mit wenig manuellem Aufwand und bereits auf geringen Datenmengen ermöglichen. Indem diese Modelle die Nutzung der Struktur von Tabellen optimieren, können realistische künstliche Daten erzeugt und die Effizienz von Datenanalysen gesteigert werden.

Das zweite Kapitel stellt **Zeitreihendaten** in den Fokus und zeigt auf, wie spezialisierte GenAI-Modelle nicht nur Prognosen erstellen, sondern auch Lücken in den Daten füllen und synthetische Daten generieren können. Diese Fähigkeiten ermöglichen eine kontextbewusste Analyse, die für Fachleute einen niedrigschwelligen Zugang zu komplexen Datenanalysen von Zeitreihen in unterschiedlichsten Branchen schafft.

Das dritte Kapitel untersucht die Vorteile von **Graphen** bei der Analyse von komplexen Netzwerkdaten. Hier demonstrieren insbesondere neue graphbasierte Transformer-Modelle und Graph Neural Networks (GNNs) als mächtige Werkzeuge ihre Stärke: sie sind in der Lage, die Struktur und Dynamik von Graphen zu erfassen. Auf diese Weise können wertvolle Einsichten in Netzwerkdaten gewonnen werden, sodass Unternehmen relevante Muster erkennen und darauf basierend bessere Entscheidungen treffen können.

Das vierte Kapitel erläutert, wie **Digitale Zwillinge** als virtuelle Repräsentationen physischer Systeme genutzt werden, um deren Zustand kontinuierlich zu überwachen und Anomalien frühzeitig zu erkennen. Durch den innovativen Einsatz von GenAI wird die Erstellung und Aktualisierung dieser Digitalen Zwillinge automatisiert. Das erleichtert den Zugang zu wertvollen Einsichten für Unternehmen erheblich und ermöglicht viele Anwendungen wie Monitoring, Betriebsoptimierung und proaktive Wartungsstrategien.

Das fünfte Kapitel beleuchtet den Einsatz **multimodaler GenAI-Modelle** in der Robotik, insbesondere bei der Steuerung von Maschinen und Robotern in dynamischen Umgebungen. Diese Modelle ermöglichen es Robotern, natürliche Sprache mit multimodalen Daten zur Steuerung und Umgebungserfassung zu verknüpfen, wodurch sie komplexe Aufgaben selbstständig und effizient durchführen können, ohne dass umfangreiche manuelle Programmierungen erforderlich sind.

Technische Grundlagen und wichtige Begriffe

GenAI und Transformer-Modelle
Die zentrale Innovation der generativen KI ist ihre Fähigkeit, Daten nicht nur zu analysieren, sondern im Rahmen unterschiedlicher Aufgabenstellungen im Kontext zu verarbeiten und kreativ zu generieren. Die ersten GenAI-Modelle bezeichnete man als große Sprachmodelle (Large Language Models, LLM), weil sie ausschließlich Text (englisch: language) verarbeiten konnten. Große Sprachmodelle können Texte fortsetzen, zusammenfassen, übersetzen oder Antworten auf Fragen geben.

Um Texte durch ein endliches Vokabular darzustellen, werden Wörter zunächst in kleinere Schnipsel (Token) zerlegt. Sie bestehen aus häufig vorkommenden Wörtern, Silben und Buchstaben und bieten den Vorteil, mit einem begrenzten

Satz von Token beliebige Wörter repräsentieren zu können. Anschließend wird jedes Token in eine Reihe von Zahlen, sogenannte Vektoren (Einbettungen, englisch: Embeddings) umgewandelt (kodiert). Die resultierende Sequenz von Embeddings dient als Input für die Transformer-Architektur (oder auch: das Transformer-Modell), das diese Einbettungen weiterverarbeitet. Das Modell selbst besteht aus vielen aufeinandergestapelten Transformer-Blöcken. Ein Transformer-Block besteht aus einem Self-Attention-Modul, und einer vollverbundenen Schicht, siehe Abbildung 1. Die vollverbundene Schicht sorgt für eine weitere Optimierung der Eingaben zur Weitergabe.

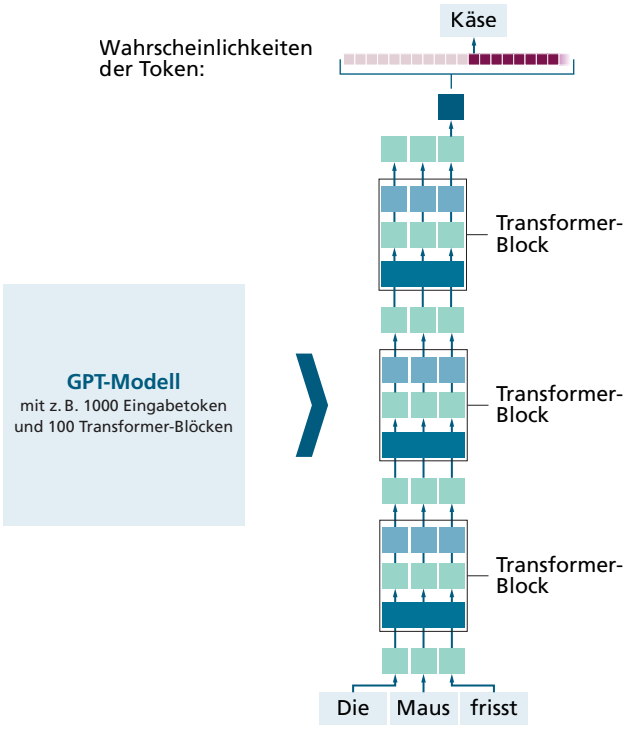


Abbildung 1: Schema eines GPT-Models, hier mit drei gestapelten Transformer-Blöcken. Blau ist die Berechnungsschicht, grün stellt die Einbettung dar und lila die Wahrscheinlichkeit. Quelle: Fraunhofer IAIS

Self-Attention – und warum sie wichtig ist

Um eine Eingabe, zum Beispiel einen Text, vollständig zu erfassen, muss das Modell die Bedeutung der einzelnen Tokens kennen. Das leistet der zentrale Mechanismus der Transformer-Architektur, die Self-Attention oder »Selbstaufmerksamkeit« [2]. Sie ermöglicht es, die Bedeutung jedes Tokens (z. B. eines Wortes in einem Satz) im Kontext aller anderen Token derselben Eingabe zu verstehen. Dies geschieht durch die Berechnung von Ähnlichkeitswerten (z. B. durch ein Skalarprodukt) zwischen den Einbettungen der Token.

Stellen wir uns einen Satz wie »Die Maus frisst Käse« vor. Für Menschen ist aufgrund des Wortes »frisst« klar, dass die »Maus« ein Tier und keine Computermaus ist. Ältere Modelle können solche kontextuellen Bedeutungen nur unzureichend erfassen, insbesondere wenn die Wörter weit voneinander entfernt stehen. Self-Attention löst das Problem der Kontextabhängigkeit, indem es für jedes Wort zwei Faktoren berücksichtigt:

Zunächst wird ein Messwert für die Ähnlichkeit der Einbettung des Referenzworts, z. B. »Maus«, zu den Einbettungen aller anderen Wörtern in der Eingabe berechnet, der auch Attention Score genannt wird. In unserem Beispiel in Abbildung 2 ist der Attention Score zwischen »Maus« und »frisst« 0,3. Anschließend wird ein neuer Einbettungsvektor für das Referenzwort »Maus« als Summe der Einbettungen aller anderen Wörter gewichtet mit deren Attention Score gebildet. Hierdurch wird der Einbettungsvektor von »Maus« mit der hoch gewichteten Einbettung von »frisst« überlagert und so die Bedeutung von Maus als Tier präzisiert.

Die einzelnen Token einer Eingabe richten also ihre Aufmerksamkeit (Attention) auf die ähnlichsten Token der Eingabesequenz und präzisieren ihre Bedeutung mithilfe der Einbettungen dieser ähnlichen Token. Die ganze Berechnung wird daher Self Attention genannt und wurde erstmals bei den eingangs erwähnten Transformer-Modellen verwendet [3]. Sie ermöglicht Transformer-Modellen, die Einbettungen zu verbessern, die schließlich die Bedeutung der Eingabetoken charakterisieren.

Um die Position jedes Tokens zu charakterisieren, wird zu jedem Eingabetoken noch eine Positionseinbettung addiert. Sowohl die Self-Attention als auch die neuen Einbettungen können für alle Token parallel berechnet werden. Diese Fähigkeit macht Transformer-Modelle extrem schnell und effizient, insbesondere bei der Verarbeitung von Kontexten mit Zigtausenden von Token. Entscheidend ist, dass viele Self-Attention Module übereinandergestapelt werden, die jeweils die Einbettungen der vorherigen Ebenen als Input erhalten. Auf diese Weise verbessern sie sukzessive die Qualität der Einbettungen und damit die Bedeutungsrepräsentation für die Token der Eingabesequenz.

Diese Fähigkeit macht Transformer-Modelle besonders geeignet für den Einsatz in Chatbots, die in Echtzeit auf Benutzeranfragen reagieren und kontextbezogene, natürliche Dialoge führen können.

Das Training von GenAI-Modellen

Ein Transformer-Modell enthält neben den Einbettungen (Tokeneinbettungen und Positionseinbettungen) weitere Parameter, die mit Zufallszahlen initialisiert und beim Training erlernt und kontinuierlich angepasst werden. Die Parameter sind das »Gedächtnis« und die Stellschrauben des Modells. Sie werden im Training so lange angepasst, bis das Modell absichtlich weggelassene (maskierte) Token möglichst gut prognostizieren kann.

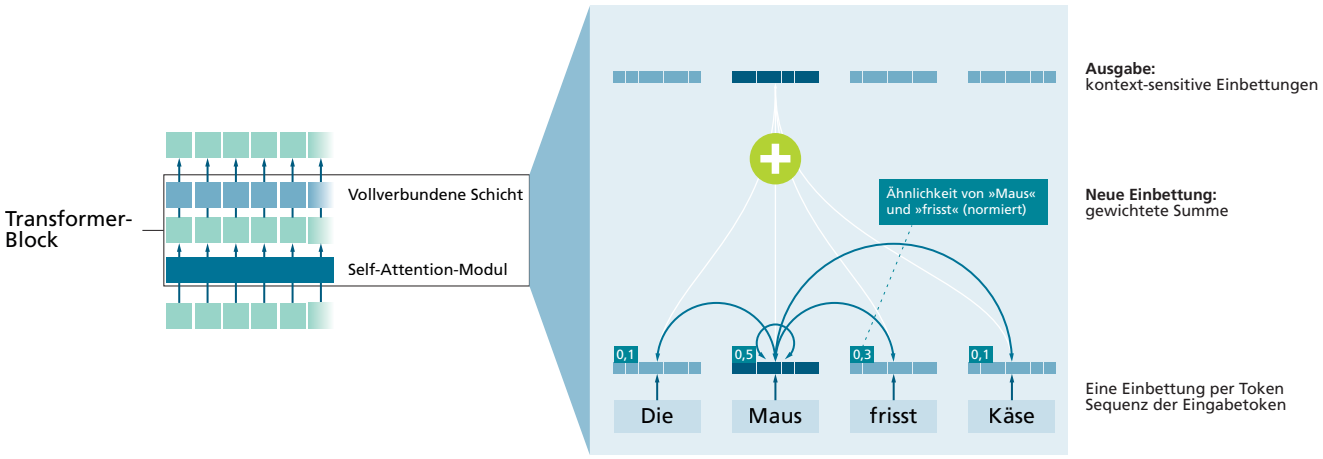


Abbildung 2: Der Transformer-Block besteht aus einem »Self-Attention-Modul«, das jedes Token mit allen anderen Token vergleicht. Quelle: Fraunhofer IAIS

Als Trainingsdaten dienen große Mengen von Text (aus dem Internet und weiteren Textsammlungen).

Ein Teil des ursprünglichen Transformer-Modells hat das Ziel, für einen Teilsatz das nächste Token zu prognostizieren. Mithilfe von Self-Attention werden zielgenaue Einbettungen für die Token des Teilsatzes berechnet, um damit die Wahrscheinlichkeiten für mögliche Token an der nächsten Position zu prognostizieren. Alle Modellparameter werden dabei durch Optimierungsverfahren so angepasst, dass die Wahrscheinlichkeit des tatsächlichen (»richtigen«) nächsten Tokens immer höher wird. Dies wird für alle Token der Trainingstexte wiederholt. Durch die Kombination von großen Mengen an Trainingsdaten, der Fähigkeit, Ähnlichkeiten von Token zu erlernen und der Optimierung der Vorhersagegenauigkeit entsteht ein Modell, das komplexe sprachliche Muster verstehen und effektiv verarbeiten kann. Diese Art von Modellen wird auch Generative Pre-trained Transformer (GPT) [4] genannt, und ist die Basis der generativen KI (GenAI).

Die Performanz von Modellen wird mit Benchmarks gemessen. Benchmarks sind standardisierte Tests, die verwendet werden, um die Vorhersagegenauigkeit oder Leistung dieser Modelle miteinander zu vergleichen.

Nutzung von GenAI-Modellen

Bei der Anwendung eines trainierten GenAI-Modells wird ein Eingabetext eingegeben, und das Modell berechnet die Wahrscheinlichkeiten für das nächste Token. Ein Token wird entsprechend dieser Wahrscheinlichkeit zufällig ausgewählt und dem Eingabetext hinzugefügt (vgl. Abbildung 1). Diese erweiterte Sequenz wird erneut ins Modell eingegeben, um das nächste Token vorherzusagen. Durch die wiederholte Vorhersage und Erweiterung der Token-Sequenz entsteht schließlich die vollständige Antwort bzw. der generierte Text. Das Modell baut die Antwort Schritt für Schritt auf, indem es jedes Mal das nächste Token vorhersagt und hinzufügt, bis die gewünschte Länge oder ein Endpunkt erreicht ist [5].

Die Leistungsfähigkeit der Modelle wächst zuverlässig mit der Anzahl ihrer Parameter und dem Umfang ihrer Trainingsdaten. Daher werden Modelle heute mit Hunderten Milliarden Parametern auf Trainingsdaten mit Billionen Token aufwendig vortrainiert. Die gewaltigen Trainingsdatenmengen machen große Sprachmodelle zu wahren Alleskönnern, die eine Vielzahl verschiedenartiger Aufgaben erfüllen können, mit denen sie im Training gar nicht direkt konfrontiert wurden. Große Modelle sind allerdings nicht immer erforderlich, da angepasste und spezialisierte Modelle mit wenigen Milliarden Parametern oft ausreichen, um hinreichend gute Ergebnisse zu erzielen.

Prompting

Überraschend ist, dass GenAI-Modelle viele Aufgaben lösen können, ohne vorher dafür speziell trainiert worden zu sein. Oft muss dem Modell nur ein Eingabetext (Prompt) gegeben werden, der die Aufgabe beschreibt, beispielsweise: »Fasse den folgenden Text mit wenigen Sätzen zusammen.« Die Qualität des Ergebnisses hängt davon ab, wie geschickt das Modell geprompted wurde, also wie genau ihm gesagt wurde, welche Ausgabe unter welchen Bedingungen es erzeugen soll. Es gibt verschiedene Methoden, einen Prompt zu verbessern. So lassen sich die Ergebnisse beispielsweise durch wenige Beispielfragen und Antworten verbessern (Few-Shot Prompts). Das Modell erfasst die Beispiele und kann sie analog zur Beantwortung der aktuellen Anfrage nutzen. Dieses »Kontextgedächtnis« ist allerdings nur temporär und das Modell merkt sich nichts dauerhaft. Ein anderer Ansatz ist das Chain-of-Thought Prompting. Dabei wird das Modell dazu angeregt, ein Problem in mehreren Schritten zu lösen und seine »Gedankengänge« und Argumentationsschritte darzulegen. Diese Technik hilft dem Modell, in mehreren Schritten zu korrekten

Schlussfolgerungen zu gelangen, und kann die Transparenz seiner Entscheidungsprozesse erhöhen.

Nachtrainieren (Fine-Tuning)
Wird eine noch höhere Zuverlässigkeit und Spezifizierung benötigt, so können Modelle durch Nachtrainieren (Fine-Tuning) für spezielle Aufgaben oder auf spezielle Domänen angepasst werden, wie beispielsweise medizinisches Vokabular oder juristische Argumentationen. Beim Nachtrainieren wird ein bereits vortrainiertes Modell mit spezifischen Daten weitertrainiert, um seine Parameter für den speziellen Kontext zu verfeinern und zu optimieren. Durch diese Anpassung der Parameter kann sich das Modell besser an die besonderen Anforderungen und Nuancen der Zielaufgabe anpassen, ohne das umfangreiche Wissen aus dem Vortraining zu vergessen. Durch das Nachtrainieren wird das Modell nicht nur spezialisierter, sondern auch zuverlässiger in der Verarbeitung der spezifischen Inhalte. Der Umrüst- und Trainingsaufwand beim Nachtrainieren ist weitaus geringer als beim vollständigen Training eines Modells von Grund auf.

LoRA (Low-Rank Adaptation) ist eine Technik zum effizienten Nachtrainieren von großen Sprachmodellen. Anstatt das gesamte Modell zu verändern, werden nur bestimmte Teile des Modells angepasst. Diese Methode reduziert den Rechenaufwand und die Speicheranforderungen erheblich, da nur ein kleiner Teil der Parameter verändert wird. LoRa ermöglicht es, Modelle schnell und kostengünstig auf neue Aufgaben oder Daten anzupassen, ohne das gesamte Modell neu trainieren zu müssen.

Falls für das Nachtrainieren keine ausreichende Menge an Trainingsdaten vorliegen oder aus Datenschutzgründen nicht verwendet werden können, kann Datenaugmentation helfen. Dabei wird die Anzahl der Trainingsdaten künstlich vergrößert, entweder durch die Modifikation vorhandener Daten, wie das Drehen von Bildern oder das Vertauschen von Datenpunkten, oder durch Erzeugung ganz neuer, synthetischer Daten [6]. Synthetische Daten sind vollständig künstlich erzeugte Daten ohne direkten Bezug zu realen Beispielen. Sie können auch von GenAI-Modellen selbst erzeugt werden.

Andere Methoden zur Spezialisierung und Optimierung
Um die Modelle noch besser auf menschliche Vorstellungen abzustimmen, bewerten Nutzerinnen und Nutzer die unterschiedlichen Antworten des Modells auf eine Vielzahl von Prompts als besser oder schlechter. Mithilfe spezieller Optimierungsverfahren (Reinforcement-Lernen) können die Modelle anschließend so angepasst werden, dass sie die bevorzugten Antworten erzeugen. Dieses Vorgehen wird als Alignment bezeichnet.

Retrieval Augmented Generation (RAG) bezeichnet eine Technik, bei der ein Sprachmodell zusätzliche Informationen aus Quellen wie Datenbanken, Dokumentensammlungen oder dem Internet erhält. Dokumente werden durch Einbettungen charakterisiert (z. B. die Summe aller Tokeneinbettungen) und in sogenannten Vektordatenbanken als Einbettungen gespeichert. Für Textstücke aus dem Prompt kann man zunächst mithilfe der Einbettungen nach ähnlichen Dokumenten in einer solchen Datenbank suchen und diese dann in den Prompt einfügen. Die Antworten werden damit aktueller und zuverlässiger, da das Modell nicht nur auf das beim Vortraining erworbene Wissen, sondern auch auf die neuesten und relevanten Informationen aus den externen Quellen zurückgreift. Eine besonders sinnvolle Quelle sind sogenannte Wissensgraphen (KG). Sie stellen strukturierte Informationen und Beziehungen zwischen Entitäten bereit, die das Modell gezielt abrufen und in die Antwortgenerierung einfließen lassen kann.

All diese Spezialisierungen und Optimierungsmethoden helfen zudem, Herausforderungen wie Halluzinationen oder Bias/Diskriminierung zu entschärfen, indem sie die Genauigkeit und Fairness der Modelle steigern.

Andere Modalitäten
Schon früh stellte sich heraus, dass GenAI nicht nur auf Text, sondern auf eine Vielzahl anderer Modalitäten (Datentypen) angewendet werden kann, wie Bilder, Musik, gesprochene Sprache oder sogar numerische Daten. Auch hier werden die Eingaben zunächst in geeignete Einbettungen transformiert, damit die spezifischen Merkmale der entsprechenden Modalität berücksichtigt und integriert werden können.
Bei Bildern werden Einbettungsvektoren beispielsweise aus kleinen Pixelbereichen erzeugt. Solche modalitätsspezifischen Embeddings werden zu einer Eingabesequenz kombiniert, wobei wiederum zusätzliche Positions-Embeddings die räumlichen oder zeitlichen Beziehungen kodieren. Auch hier wird die Transformer-Architektur angewendet, um das nächste Token einer Sequenz vorherzusagen. Die Self-Attention setzt die Token der Sequenz zueinander in Beziehung, wodurch die berechneten Einbettungen die Bedeutung der jeweiligen Elemente (Worte, Bildteile, Audioschnipsel, ...) präzise erfassen können.

Neben den großen Sprachmodellen gibt es andere spezialisierte Transformer-Modelle, die auf eine jeweilige Modalität angepasst sind. Das sind zum Beispiel Large Time Series Models (LTM) für die Verarbeitung von numerischen Daten mit einem Zeitbezug, Vision Language Models (VLM) für die Integration von visuellen und sprachlichen Informationen oder Graph Neural Networks (GNN) für die Analyse von Datenstrukturen in Form von Graphen. Diese spezialisierten Modelle sind darauf ausgelegt, die spezifischen Anforderungen ihrer jeweiligen Modalität zu erfüllen und bieten dadurch eine optimierte Leistung für bestimmte Aufgabenbereiche.

Multimodale Transformer-Modelle
Multimodale Transformer-Modelle integrieren Informationen aus verschiedenen Modalitäten, wie Text, Bild und Audio, um komplexe Zusammenhänge auch zwischen den Modalitäten zu verstehen. Die Modelle verarbeiten alle verschiedenen Eingaben unterschiedlicher Modalität und überführen sie in geeignete numerische Embeddings, die die Bedeutungen der Elemente der unterschiedlichen Datentypen erfassen. Sie können dann Aufgaben bewältigen, die Informationen aus mehreren Modalitäten kombinieren, wie die Beschreibung von Bildern, Videos oder gesprochene Sprache. So kann das Modell beispielsweise lernen, das Wort »Hund« in einem Text mit dem Hund in einem Foto von einem Wohnzimmer zu verbinden.

Diese Modelle sind besonders wertvoll für Anwendungen, die eine umfassende Betrachtung von Daten erfordern, und eröffnen neue Möglichkeiten in der Mensch-Maschine-Interaktion und der intelligenten Datenanalyse. Allerdings sind sie auch sehr groß und rechenintensiv, da sie erhebliche Rechenleistung und Speicherplatz benötigen, um die komplexen Berechnungen für die Integration der verschiedenen Modalitäten durchzuführen.

Foundation-Modelle (Grundmodelle)
Als Foundation-Modelle [7] oder Grundmodelle werden Prognosemodelle bezeichnet, die einerseits auf einer umfangreichen Datenbasis trainiert wurden und die andererseits an eine breite Palette nachgelagerter Aufgaben angepasst werden können. Sie sind nahezu universell einsetzbar. Transformer-Modelle sind die bekannteste Art von Foundation-Modellen. Aktuelle Modelle, z. B. ChatGPT-4o oder Gemini, haben viele Hundert Milliarden Parameter und werden mit enormen Medienbeständen mit unterschiedlichen Datentypen (Texte, Bilder, Sprache) trainiert. Sie erzielen daher auch medienübergreifend eine bisher unerreichte Prognosequalität.

Während die ersten großen Sprachmodelle die vorgegebene Aufgabe in einem Schritt gelöst haben, können sogenannte Reasoning-Modelle inzwischen komplexe Probleme zerlegen und schrittweise lösen. Ausgehend von Chain-of-Thought Prompts können die Modelle Informationen logisch verknüpfen, Alternativen abwägen und fundierte Entscheidungen treffen. Dieses schrittweise Vorgehen beim »Reasoning« (nachvollziehbares Denken) ist besonders vielversprechend bei mathematischen Aufgaben oder im Ingenieurbereich und verspricht ganz neue Möglichkeiten.

Im Zentrum der Aufmerksamkeit stehen derzeit KI-Agenten. Im Kern eines KI-Agenten steht meist ein LLM, das über einen Prompt eine Aufgabe erhält. Daraufhin muss der Agent einen Lösungsweg planen, beispielsweise mithilfe von Reasoning-Techniken. Anschließend kann er verschiedene Hilfsmodule ansprechen, wie beispielsweise Websuche, externe Sensoren, oder Schnittstellen zur Steuerung eines Roboters oder eines Computers. Das Modell wurde darauf trainiert, den Modulen in einem geeigneten Format Anweisungen zu erteilen, und kann deren Ausgaben interpretieren. Über eine Speicherkomponente kann das Modell Zwischenergebnisse speichern und später darauf zugreifen. Bei der Lösung komplexer Aufgaben kann das Modell Abläufe simulieren und deren Ergebnisse abschätzen. Es durchläuft eine Abfolge von »Denken«, »Handeln« und »Beobachten« und wiederholt diese Schritte bei Bedarf, bis eine tragbare Lösung erreicht ist. KI-Agenten können komplexe Aufgaben weitgehend autonom ausführen. Sie interagieren mit Nutzenden, ihrer Umwelt und anderen Modulen und Agenten, um proaktiv Lösungen zu entwickeln und umzusetzen. KI-Agenten heben GenAI auf eine neue Stufe.

Angepasste Foundation-Modelle für wirtschaftliche und technische Datenwelten
Dieses Whitepaper konzentriert sich auf Foundation-Modelle, die auf besondere Modalitäten (Datentypen) angewendet werden. Diese stehen weniger im Zentrum des öffentlichen Interesses, gewinnen aber für wirtschaftliche Anwendungen in Unternehmen zunehmend an Bedeutung. Dazu gehören Zeitreihen, Tabellen, netzartige Graphstrukturen, Digitale Zwillinge und Robotik-Daten, siehe Abbildung 3. Für diese speziellen Datentypen müssen auch hier die vorliegenden Daten zunächst in eine Sequenz von Einbettungsvektoren umgewandelt werden. Die Kodierung muss die besonderen Eigenschaften und Beziehungen in den Daten, wie etwa zwischen Spalten, Zeilen und numerischen Werten in Tabellen oder zwischen Knoten und Kanten in Graphen optimal erfassen. Spezialmodelle nutzen manchmal abgewandelte Transformer-Architekturen oder spezielle Trainingsmethoden, um mit diesen spezifischen Datentypen besser umzugehen als ein klassisches großes Sprachmodell. Dabei können die Modelle je nach Aufgabe nur einen Datentyp oder mehrere Datentypen gleichzeitig erfassen und vorhersagen.

Diese Foundation-Modelle können auf verschiedene Weise angepasst und genutzt werden, um spezifische Anforderungen zu erfüllen:

Einerseits können sie auf einem speziellen Trainingsdatensatz, bestehend aus branchenspezifischen Daten, vortrainiert werden. Zusätzlich kann ein vortrainiertes Modell mit unternehmensinternen Daten nachtrainiert werden, um es auf die besonderen Erfordernisse der eigenen Anwendung abzustimmen. Oder man setzt Retrieval Augmented Generation (RAG) ein, um die Modelle mit zusätzlichen Informationen aus internen Unternehmensquellen zu versorgen.

Darüber hinaus gibt es mittlerweile vortrainierte Modelle für neue Modalitäten, die auf Millionen von Trainingsdaten basieren und durch geschicktes Prompting ohne

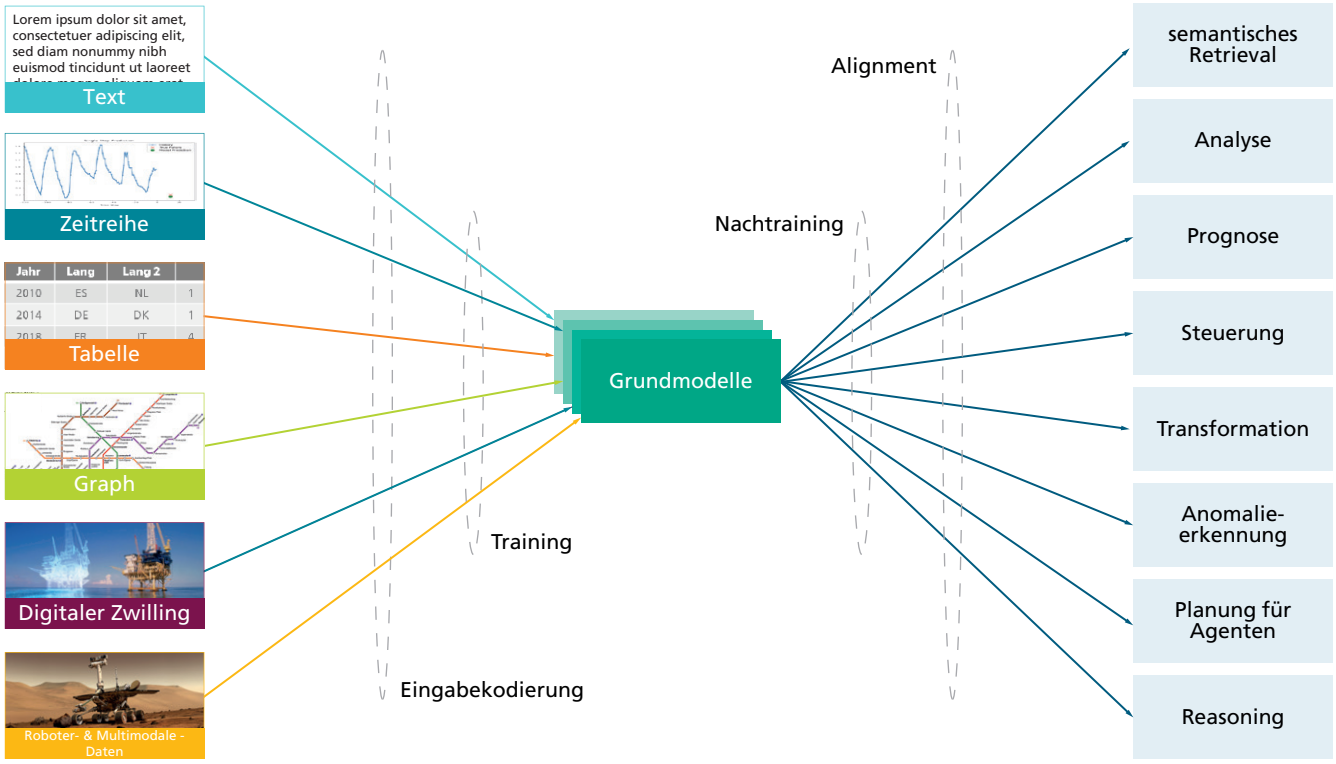


Abbildung 3: Ein Foundation-Modell erhält eine Eingabesequenz aus einer oder mehreren Medienarten. Die Eingabekodierung erzeugt Embeddings aus den Elementen der Eingabesequenz. Durch Training werden die Parameter des Grundmodells optimiert. Durch Nachtrainieren und Alignment wird es auf seine spezifischen Aufgaben vorbereitet. Aktuelle Inhalte können aus einem RAG-System durch Retrieval abgerufen werden. Über Prompting wird das Modell angewiesen, eine Aufgabe zu lösen.
Quelle: Fraunhofer IAIS

weiteres Nachtrainieren für eigene Aufgaben eingesetzt werden können, wie etwa für Zeitreihenprognosen und Tabellenverarbeitung.

Die verschiedenen Ansätze aus diesem Whitepaper zeigen anschaulich, dass die Verwendung von spezialisierten Modellen, angepassten Sprachmodellen, multimodalen Modellen oder der Kombination mehrerer spezialisierter Modelle jeweils einzigartige Vorteile bieten, um unternehmensspezifische Anforderungen und Herausforderungen zu adressieren. Die Prognoseleistungen von klassischen Modellen werden dabei in aller Regel übertroffen. Diese Vielfalt an Strategien ermöglicht es, maßgeschneiderte Lösungen zu entwickeln, die sowohl die Effizienz als auch die Genauigkeit der Anwendungen maximieren.

1. Tabellarische Daten und Matrizen

1.1. Die Datenstruktur

Tabellarische Daten und Matrizen sind überall in der digitalen Welt zu finden, von Excel-Tabellen über Datenbanken bis zu Matrizen in Simulationen und Maschinellem Lernen.

Eine Tabelle besteht aus Zeilen und Spalten, wobei jede Zeile einen vollständigen Datensatz repräsentiert und jede Spalte ein spezifisches Merkmal dieses Datensatzes beschreibt. Diese Struktur ermöglicht die Speicherung heterogener Datentypen wie Text, Zahlen und Datumsangaben, wie zum Beispiel die einfache Kundenliste in Abbildung 4, die Name, Alter und Wohnort erfasst. Diese Flexibilität macht tabellarische Daten besonders nützlich für die Datenanalyse und Berichterstattung, wo komplexe Informationen ausgewertet werden.

Wie die Gegenüberstellung in Abbildung 4 zeigt, sind Matrizen eine spezielle Form tabellarischer Daten, die sich auf numerische Werte konzentrieren. Sie werden häufig verwendet, um komplexe Datenmuster darzustellen und mathematische Operationen durchzuführen, wie zum Beispiel für Produktbewertungen, bei denen die Zeilen Produkte und die Spalten verschiedene Bewertungskriterien darstellen. Matrizen ermöglichen eine effiziente Verarbeitung großer Datenmengen, was sie für Maschinelles Lernen besonders wertvoll macht.

Die enge Verwandtschaft dieser beiden Datenstrukturen ermöglicht es, sie synergistisch zu nutzen. Unternehmen können tabellarische Daten verwenden, um eine breite Palette von Informationen zu erfassen und zu organisieren, die für die Entscheidungsfindung ausschlaggebend sind. Diese Daten können dann in Matrizen umgewandelt werden,

Name	Alter	Wohnort
Anna	35	Berlin
Ben	29	München
Claudia	22	Hamburg
David	42	Köln

Tabellarische Daten

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Matrix

Abbildung 4: Beispiel einer Tabelle und einer Matrix
Quelle: Fraunhofer IAIS

um quantitative Analysen durchzuführen und Muster oder Zusammenhänge zu identifizieren. Moderne Technologien wie Sprachmodelle (LLMs), bieten neue Möglichkeiten, tabellarische Daten und Matrizen sowohl intelligent als auch automatisiert zu verarbeiten und zu analysieren. Dadurch können Unternehmen nicht nur historische Daten auswerten, sondern auch Echtzeitanalysen durchführen und proaktive Entscheidungen treffen.

1.2. Warum GenAI für diese Datenstruktur?

Strukturierte, insbesondere tabellarische Daten sind von zentraler Bedeutung im klassischen Maschinellen Lernen. Viele Branchen wie Finanzen, Medizin, Handel, Landwirtschaft und Bildung setzen stark auf relationale Datenbanken und besitzen damit umfangreiche Trainingsdaten für klassische KI-Methoden wie Entscheidungsbäume und Gradient Boosting [8]. Gradient Boosting nutzt viele einfache Entscheidungsbäume, um in Kombination eine möglichst genaue Vorhersage zu treffen.

Dass dieses Verfahren noch immer Stand der Technik ist, liegt unter anderem an den besonderen Herausforderungen, die sich vor allem aus der Heterogenität, Qualität und Quantität der enthaltenen Informationen ergeben [9]. Während Matrizen nur aus numerischen Werten bestehen, umfassen tabellarische Daten eine Mischung aus Zahlenwerten in Spalten wie Alter, Preis oder Gewicht und kategorischen Werten in Spalten wie Land, Farbe oder Geschlecht. In beiden Fällen besteht eine zusätzliche Herausforderung darin, die Beziehung zwischen den enthaltenen Werten herzustellen und somit die durch die Struktur gegebenen Informationen optimal zu nutzen. Weiterhin sind reale Datensätze häufig von fehlenden Werten geprägt, können eine große Menge an Spalten, aber vergleichsweise wenige Zeilen aufweisen oder stark unausgewogene Verteilungen beinhalten.

GenAI, mit seinen Stärken auf unstrukturierten Daten wie Text, Bild und sogar Audio, besitzt das Potenzial, auch diese besonderen Herausforderungen bei Tabellen auf innovative Weise zu adressieren. Ein anschauliches Beispiel dafür ist der medizinische Bereich, in dem enorme Mengen strukturierter Daten anfallen – von elektronischen Patientenakten über klinische Studien bis hin zu Verwaltungsdaten.

Ein zentrales Problem in diesem Bereich sind unausgewogene Daten. Bei seltenen Krankheiten ist die Menge an verfügbaren Informationen oft sehr gering. Zudem können die existierenden Daten aus Datenschutzgründen nicht ohne Weiteres geteilt werden, was die Datenmenge für Analysen weiter reduziert. GenAI kann diese Probleme lösen, indem auf Basis bekannter Krankheitsdaten neue, plausible Patientendaten erzeugt werden. Ein Modell, das diesen Ansatz verfolgt, ist das Tabular Transformer Generative Adversarial Network (TT-GAN) [10]. Dieses Modell passt den Transformer speziell für die Verarbeitung tabellarischer Daten an, indem es jeden Wert einer Zeile unter Einbezug aller Spaltenwerte einbettet. Durch Self-Attention werden daraufhin die Beziehungen zwischen den Vektoren untersucht, die die Spaltenwerte repräsentieren. Damit wird die zweidimensionale Struktur von Tabellen explizit berücksichtigt. Nachdem das Modell auf einem Datensatz trainiert wurde, liefert es als Ergebnis künstlich generierte realistische Daten. Diese können von KI- oder GenAI-Modellen genutzt werden, um beispielsweise Krankheitsbilder besser zu erkennen. Die Generierung solcher Daten mit GenAI-Modellen wie TT-GAN löst auch das Datenschutzproblem: Da die künstlich erzeugten Daten statistisch realistisch, aber nicht auf echte Patienten zurückzuführen sind, wird der Austausch von Daten erleichtert. Dadurch können große, vielfältige Gesundheitsdatensätze entstehen, die die Entwicklung neuer Medikamente und die Qualität von KI-Modellen zur Krankheitsdiagnose verbessern können.

Die Möglichkeit zur Generierung künstlicher tabellarischer Daten mit realistischen, aber gleichzeitig vielfältigen Merkmalen hat auch die Entwicklung von tabellarischen Grundmodellen (Foundation-Modellen) unterstützt. Ein Beispiel hierfür ist das Tabular Prior-data Fitted Network (TabPFN) [11], das auf Transformer-Modellen basiert und auf einer großen Menge künstlich generierter Tabellen trainiert wurde. TabPFN legt ebenfalls ein besonderes Augenmerk auf die explizite Repräsentation der durch die Zeilen und Spalten gegebenen zweidimensionalen Struktur von Tabellen, indem die Self-Attention des Transformer-Modells auf eine zweite Dimension ausgeweitet wird. Damit werden sowohl die Beziehungen zwischen Werten innerhalb einer Zeile als auch die Verteilung innerhalb einer Spalte analysiert. Dies führt zu einem Foundation-Modell, das bereits ohne datensatzspezifisches Training die Vorhersagegenauigkeit im Vergleich zu Gradient-Boosting-Entscheidungsbäumen verbessert, während diese für jeden neuen Datensatz von Grund auf trainiert werden müssen. Auch wenn TabPFN auf Datensätze mit bis zu 10.000 Einträgen und 500 Merkmalen spezialisiert ist, zeigt es dennoch das Potenzial von tabellarischen Foundation-Modellen.

Die Fähigkeit, mithilfe von vortrainierten Foundation-Modellen ohne umfangreiche manuelle Vorarbeit präzise Vorhersagen zu treffen, stellt einen wesentlichen Fortschritt dar – besonders in Szenarien, in denen die schnelle Verarbeitung und Analyse von strukturierten Daten gefragt ist. Neben dem medizinischen Bereich bieten GenAI-Modelle auch anderen Branchen erhebliche Vorteile. Beispielsweise, wenn im Handel häufig ohne oder nur mit wenigen Informationen zu kurzlebigen saisonalen Produkten Entscheidungen getroffen werden müssen, können tabellarische Foundation-Modelle genau auf diesen Fall abgestimmt werden. Im Versicherungswesen können solche Modelle in der Erstellung von Risikoprofilen anhand von Kundendaten, Regionen und Schadenshistorie unterstützen. Oder es können im Finanzwesen auf Basis weniger Daten wahrscheinliche Betrugsversuche erkannt werden.

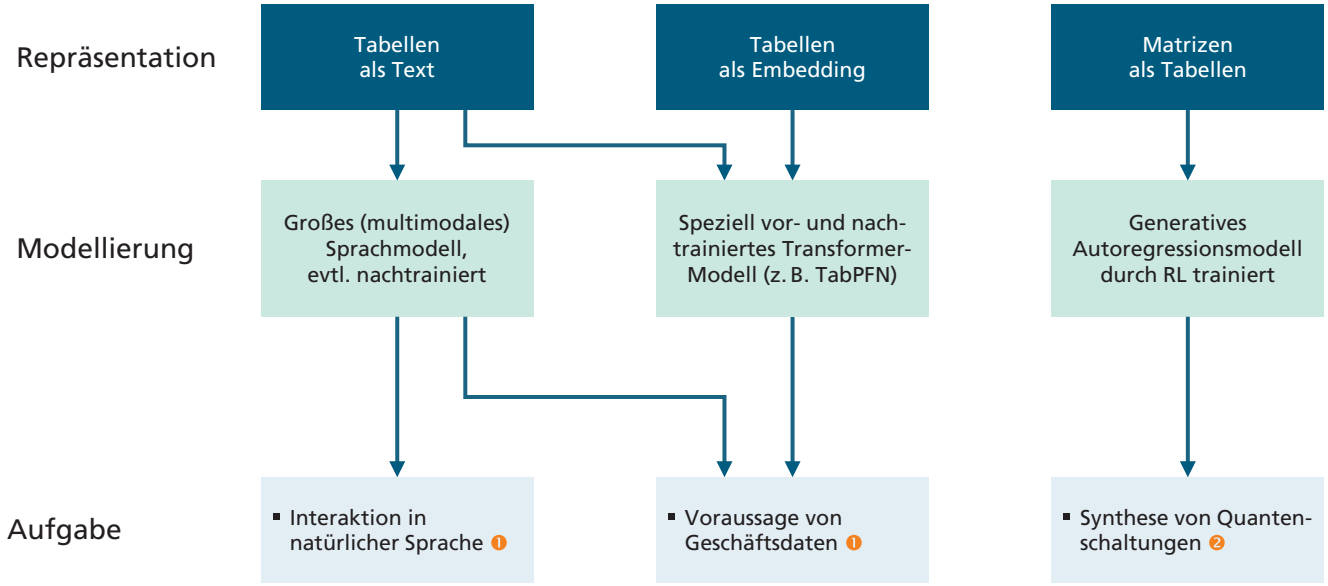


Abbildung 5: GenAI mit Tabellen & Matrizen
Quelle: Fraunhofer IAIS

Auch wenn die tabellarischen Foundation-Modelle noch nicht so ausgereift sind wie die Grundmodelle auf Text oder Bild, markieren die aktuellen Forschungsergebnisse einen entscheidenden Schritt hin zu einer datengetriebenen Zukunft, in der GenAI eine immer zentralere Rolle in der Entscheidungsfindung und Innovation für Unternehmen und Forschungseinrichtungen spielen wird.

Abbildung 5 bietet einen strukturierten Überblick darüber, wie Tabellen und Matrizen in den nachfolgenden Use Cases verarbeitet und in Kombination mit GenAI eingesetzt werden.

Tabelle 2: Tabellen und Matrizen – Use Cases in der Übersicht

Branche/ Bereich	Use Case	Aufgabentyp	Datentyp & Datenstruktur	Modell oder Methode
Unternehmen	❶ Vorhersage von strukturierten Geschäftsdaten für Management und Vertrieb	Geschäftliche Kennzahlen prognostizieren	Tabelle mit historischen Transaktionsdaten	Transformer-basierte tabellarische Grundmodelle (z. B. TabPFN)
Forschung & Entwicklung	❷ Synthese optimierter Quantenschaltungen für Hardwareprovider und F&E	Optimierte Quantenschaltungen synthetisieren	Strukturierte numerische Tabellendaten mit komplexwertigen unitären Matrizen	RL-basiertes Training mit autoregressiver Inferenz

1.3. Use Cases

1.3.1. Voraussage von strukturierten Geschäftsdaten ❶

Die Prognose geschäftlicher Kennzahlen ist entscheidend für die strategische Planung und Entscheidungsfindung in Bereichen wie Absatzplanung, Finanzvorhersage, Lageroptimierung und Risikobewertung. Strukturierte Tabellendaten bilden die Grundlage für fundierte Prognosen. Gradient-Boosting-Entscheidungsbäume, wie XGBoost [12], gelten als bewährter Standard. Allerdings erfordert jede neue Datenquelle oder Aktualisierung der Daten erheblichen manuellen Aufwand, u. a. für die sorgfältige Bereinigung, Formatierung und Transformation der Daten. Dies umfasst das Entfernen von Ausreißern, das Auffüllen fehlender Werte und die Umwandlung textbasierter Informationen in numerische oder kategorische Formate, die das Modell verarbeiten kann. Dieser zeitintensive Prozess verlangt ein tiefes Verständnis der Geschäftsprozesse und -kontexte, um die relevanten Informationen korrekt zu erfassen.

GenAI eröffnet neue Möglichkeiten zur Nutzung strukturierter Daten. Diese Technologie ermöglicht sprachliche Interaktionen sowie flexible und schnelle Vorhersagen oder Analysen, siehe Abbildung 6. Anders als traditionelle Ansätze sind LLMs wie GPT oder Llama in der Lage, Aufgaben zu bewältigen, die nicht explizit Teil ihres Trainings waren, sofern sie in natürlicher Sprache beschrieben und mit Beispielen bereitgestellt werden. Beispiele können in Textform wie »Name ist Max, Alter ist 31« oder als speziell für LLMs erstellte Embeddings vorliegen. Durch erlerntes Weltwissen und die Fähigkeit zur Mustererkennung liefert GenAI als Antwort fundierte Hypothesen. Dies ermöglicht schnellere Prognosen ohne umfangreiche manuelle Vorarbeit, denn textbasierte Informationen werden dank des Sprachverständnisses der Modelle aktiv genutzt, wodurch der vorhandene Informationsgehalt effizienter ausgeschöpft wird.

In einem Projekt des Fraunhofer IAIS wurde die Genauigkeit verschiedener GenAI-Modelle untersucht. Llama-3 zeigte mit wenigen Beispielen bessere Ergebnisse als einfache Heuristiken, erreichte jedoch nicht die Leistung von XGBoost. Das Foundation-Modell Tabula [13] erzielte auf Datensätzen mit hohem Textanteil vergleichbare Ergebnisse zu traditionellen KI-Modellen. Bei überwiegend numerischen Spalten war zusätzliches Nachtrainieren erforderlich, doch damit konnten ähnliche Ergebnisse erzielen werden. Tabula kann aufgrund

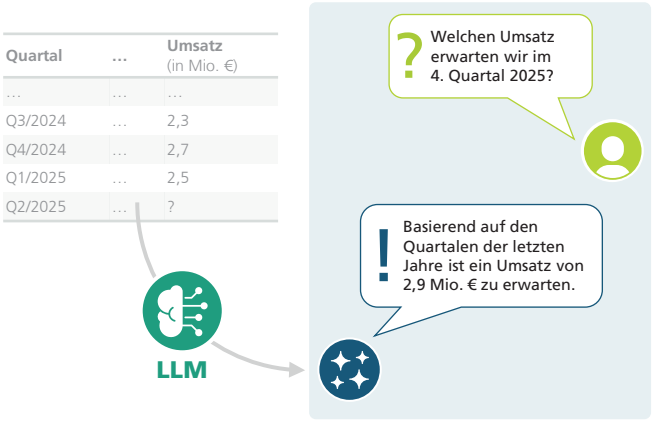


Abbildung 6: Vereinfachte Darstellung der Voraussage mit LLMs
Quelle: Fraunhofer IAIS

textbasierter Datenverarbeitung nur wenige Beispiele verarbeiten. Dagegen können tabellarische Foundation-Modelle wie TabPFN und TabDPT [9] größere Mengen an Beispielen nutzen, da sie Daten als Embeddings verarbeiten. Diese Modelle erzielten in allen Experimenten ohne Nachtrainieren ähnlich gute oder bessere Ergebnisse als XGBoost.

GenAI eröffnet neue Möglichkeiten für Prognosen von Geschäftsdaten in verschiedenen Kontexten. Für fest definierte Voraussagen auf sich selten ändernden Daten sind XGBoost und ähnliche Methoden weiterhin geeignet. Für schnelle oder interaktive Anfragen auf häufig aktualisierten Datensätzen kann GenAI eine neue starke Alternative bieten, insbesondere wenn Textinformationen zentraler Bestandteil der Daten sind. Die LLMs könnten dabei die Schnittstelle in natürlicher Sprache darstellen und Anfragen mithilfe von Foundation-Modellen beantworten.

Kontakt

Dr. Tim Wirtz
Fraunhofer IAIS
tim.wirtz@iais.fraunhofer.de

1.3.2. Synthetisierung von Quantenschaltungen ❷

Quantencomputing hat sich in den letzten Jahren als ein vielversprechendes und revolutionäres Feld der Informatik etabliert. Es nutzt die Prinzipien der Quantenmechanik, um potenziell schnellere und effizientere Algorithmen für eine Vielzahl von Anwendungen zu entwickeln, darunter kombinatorische Optimierung und chemische Simulationen. Kürzlich erzielte experimentelle Durchbrüche in diesem Forschungsfeld erhöhen den Bedarf an optimierten Quantenschaltungen, die die fundamentalen Recheneinheiten eines Quantencomputers, die sogenannten Qubits, nutzen.

Eine Quantenschaltung ist eine Abfolge von Quantengattern, die auf spezifischen Quantenhardwaretypen ausgeführt werden können. Jeder Hardwaretyp hat seine eigenen nativen Gatter. Die zentrale Herausforderung im Kompilierungsprozess ist die effiziente Überführung einer abstrakten Quantenschaltung in native Gatter mit minimalem Einsatz kostenintensiver Gatter. Letztere sind hinsichtlich ihrer Ausführungszeit, Fehleranfälligkeit oder dem Aufwand für eine fehlertolerante Implementierung besonders problematisch. Jede Quantenschaltung und jedes Gatter lässt sich als unitäre Matrix darstellen, deren Dimension exponentiell mit der Anzahl der Qubits wächst und somit schnell unhandhabbar wird. Dennoch kann es oft sinnvoll sein, Subschaltungen mit wenigen Qubits zunächst als Matrix zu betrachten und in eine Gattersequenz zu überführen. Die Problemstellung der Synthese besteht darin, für eine gegebene unitäre Matrix eine Gattersequenz zu finden, die diese Matrix möglichst gut approximiert und gleichzeitig eine minimale Anzahl bestimmter Gattertypen enthält.

Verschiedene Lösungsansätze sind bekannt, diese sind jedoch entweder durch hohe Laufzeiten oder suboptimale Lösungen limitiert. Aufgrund des strukturierten Suchraums bietet sich für die Synthese solcher Quantenschaltungen Reinforcement Learning (RL) an, siehe Abbildung 7. RL ermöglicht es durch Trial-and-Error, basierend auf einem Belohnungssignal, optimierte Lösungsstrategien zu entwickeln. Während des Trainingsprozesses identifiziert der Algorithmus unbekannte hilfreiche Heuristiken und lernt, diese zur Lösung des Problems auszunutzen. Dadurch entsteht ein generatives Modell, das Schaltkreise autoregressiv Gatter für Gatter erzeugt.

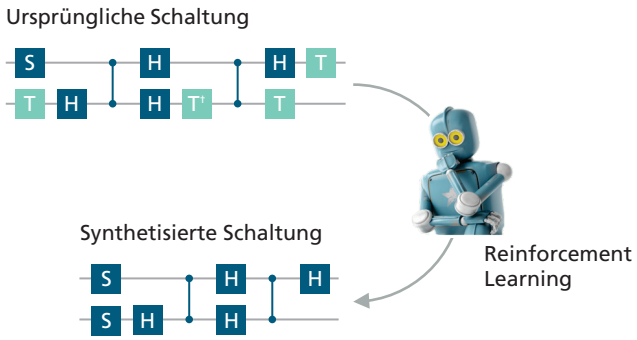


Abbildung 7: Die Abbildung zeigt zwei Quantenschaltungen, bei denen ein RL-Algorithmus die ursprüngliche Schaltung als unitäre Matrix optimiert, um die Synthese ohne T-Gatter zu erreichen. Quelle: Fraunhofer IIS

Der am Fraunhofer IIS entwickelte RL-Algorithmus erhält die zu synthetisierende Matrix als Eingabe in tabellarischer Form. Zu Beginn entspricht das Syntheseergebnis der Identitätsmatrix, da zunächst noch keine Gatter vorhanden sind. Schrittweise platziert der Algorithmus das nächste native Gatter, bis die Synthese erfolgreich abgeschlossen ist, was durch ein Belohnungssignal bestätigt wird. Bei jedem Platzieren eines kostenintensiven Gatters wird der Algorithmus bestraft.

Für das Training wurden zufällig generierte Schaltungen verwendet. Nach Abschluss des Trainings kann der Algorithmus auf beliebige neue Schaltungen angewendet werden. Die vielversprechenden Ergebnisse sind für zukünftige fehlertolerante Implementierungen von Quantenschaltungen von enormer Bedeutung und bieten eine solide Grundlage für weitere Optimierungen und Verbesserungen in diesem spannenden Forschungsbereich.

Kontakt

Dr. Daniel Scherer
Fraunhofer IIS
daniel.scherer2@iis.fraunhofer.de

2. Zeitreihen

2.1. Die Datenstruktur

Eine Zeitreihenanalyse ist entscheidend für die Entwicklung eines Verständnisses messbarer Veränderungen über die Zeit. Dies macht sie in vielen Bereichen unverzichtbar, in denen zeitlich-induzierte Dynamiken und zugrunde liegende Muster strategische Entscheidungen ermöglichen (siehe Abbildung 8). Beispiele für Zeitreihendaten reichen von einfachen Sensormessungen über Bewegungstrajektorien bis hin zu Videodaten und Prozessabläufen.

Die Datenpunkte umfassen den Zeitstempel, der die Position jedes Datenpunkts in der Zeitreihe bestimmt, sowie einen oder mehrere zugehörige Messwerte. Die Messwerte können hierbei numerisch sein, wie etwa bei Temperaturmessungen und Verkaufszahlen oder kategorisch, wie im Fall von Statusmeldungen. Ein weiteres charakteristisches Merkmal ist die zeitliche Auflösung, mit der die Datenpunkte erfasst werden. Unterschiedliche Auflösungen sind dabei oftmals abhängig vom jeweiligen Analysezweck und Anwendungsgebiet.

Die Anwendungsgebiete für Zeitreihendaten sind vielfältig und reichen von der Analyse von Aktienkursen und Marktindizes im Finanzwesen zur Optimierung von Investitionsstrategien bis hin zur Überwachung von Patientenparametern wie Herzfrequenz oder Blutdruck im Gesundheitswesen, um frühzeitig auf gesundheitliche Veränderungen reagieren zu können. In der Produktion werden Maschinenleistung, Produktionsausfallzeiten und sonstige Ereignisse erfasst, um Wartungszyklen und Prozessoptimierung zu planen. Wegbereiter hierfür ist die weite Verbreitung von IoT-Technologien. Allein für Innovationen

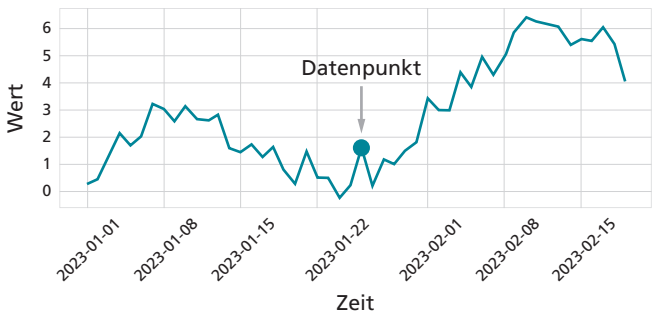


Abbildung 8: Die Datenstruktur Zeitreihen
Quelle: Fraunhofer FIT

durch Zeitreihenprognosen beläuft sich der für das Jahr 2024 geschätzte, weltweite Marktwert auf 0,31 Mrd. US-Dollar, bis 2033 soll er auf 0,47 Mrd. US-Dollar anwachsen [14].

Zeitreihendaten bieten ein enormes Potenzial – vorausgesetzt, sie werden kontextbewusst modelliert und analysiert.

Zusammengefasst bieten Zeitreihendaten und deren Analysen Unternehmen zahlreiche Anknüpfungspunkte, um historische Trends zu identifizieren, zukünftige Entwicklungen zu prognostizieren und operative Entscheidungen zu optimieren.

2.2. Warum GenAI für diese Datenstruktur?

Traditionell wurden Zeitreihen mit klassischen, oft linearen Methoden wie ARIMA analysiert. Die neuen Transformer-Modelle können Beziehungen zwischen Datenpunkten unabhängig davon berücksichtigen, wie weit diese in der Zeitreihe voneinander entfernt sind. Dabei kodieren die Token statt der Wörter die Messwerte zu den verschiedenen Zeitpunkten der Eingabe. Solche Ansätze sind besonders nützlich für Finanzzeitreihendaten, bei denen vergangene Ereignisse einen erheblichen Einfluss auf zukünftige Preise haben können. Da Transformer ihre Eingabesequenz parallel verarbeiten, sind ihre Berechnungen viel schneller und besser skalierbar. Zudem können Korrelationen zwischen mehreren Zeitreihen effizient erfasst werden. In einer großen Zahl von Benchmarks haben Transformer-Modelle im Vergleich zu klassischen Ansätzen zu genaueren Prognosen geführt.

Neben der automatisierten Datenanalyse und präzisen Prognose eignen sich GenAI-Modelle auch für neue Aufgaben wie das Füllen von Lücken in den Daten, die Generierung abgeleiteter Merkmale und synthetischer Daten, die Fortsetzung der Zeitreihe unter verschiedenen Szenarien und die individuelle Berichterstellung. Während Zeitreihenanalysen früher meist spezialisierten Datenexpertinnen und Datenexperten vorbehalten waren, ermöglicht der Einsatz von GenAI heute einen niedrigschwelligen Zugang – unabhängig vom individuellen Analysewissen.

Wie in Abbildung 9 dargestellt, gibt es verschiedene Möglichkeiten, GenAI-Modelle für Zeitreihendaten zu nutzen. Vortrainierte Sprachmodelle wie GPT-4, Claude oder LLaMA

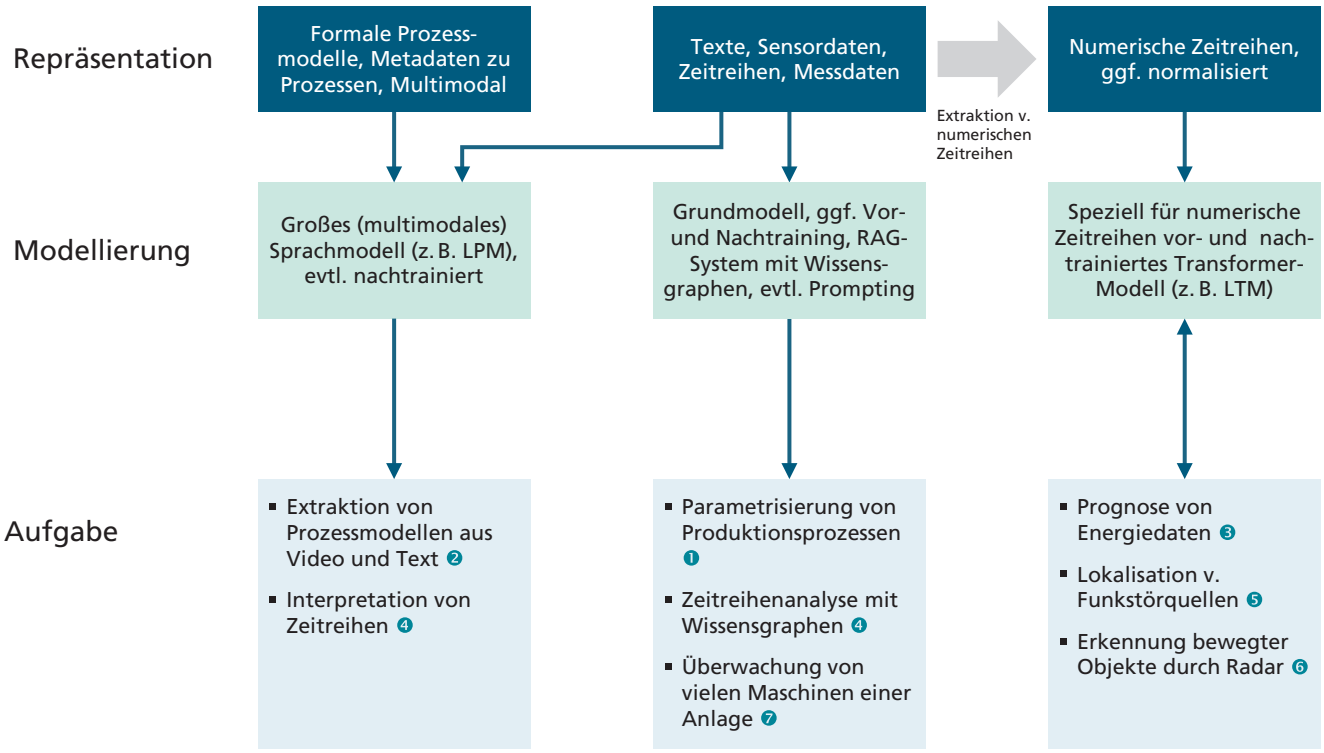


Abbildung 9: Strukturierter Überblick, wie Zeitreihen in den nachfolgenden Use Cases verarbeitet und GenAI eingesetzt werden kann. Quelle: Fraunhofer FIT

können Zeitreihendaten in textueller Form interpretieren und mit numerischen sowie semantischen Informationen anreichern. Durch gezielte Prompt-Entwicklung lassen sich diese Modelle direkt für Aufgaben wie Klassifikation, Prognose oder Anomalieerkennung einsetzen – ohne dass ein aufwendiges Modelltraining erforderlich wird. Das sogenannte Zero-Shot ermöglicht einen schnellen Einstieg in die Analyse, auch bei begrenzter Datenverfügbarkeit.

Darüber hinaus sind LLMs in der Lage, numerische Werte im Kontext zu interpretieren. So kann ein Modell beispielsweise den Wert »90« in einer Spalte »Gewicht« korrekt als Körpergewicht einer erwachsenen Person einordnen – unter Einbezug seines umfassenden Weltwissens über typische Werteberei-che und Zusammenhänge. Dieses semantische Embedding numerischer Daten ist ein entscheidender Vorteil gegenüber klassischen Verfahren.

Für spezifischere Anwendungsfälle können vortrainierte Modelle durch Nachtrainieren (Fine-Tuning) weiter spezialisiert werden. Dies erlaubt es, unternehmensspezifisches Wissen oder branchenspezifische Besonderheiten gezielt zu integrieren. Alternativ lassen sich LLMs in sogenannte Retrieval-Augmented-Generation (RAG)-Architekturen einbetten, bei denen externe Wissensquellen wie Datenbanken oder Wissensgraphen eingebunden werden, um kontextbezogene Antworten zu ermöglichen.

Ein weiterer Ansatz besteht in der Entwicklung und dem Training spezialisierter Transformer-Modelle direkt auf Zeitreihendaten – sogenannter Large Timeseries Models (LTMs). Diese Modelle nutzen speziell entwickelte Embeddings, die besser auf die Struktur und Dynamik von Zeitreihen abgestimmt sind als reine Textrepräsentationen. Auch LTMs lassen sich durch Nachtrainieren weiter anpassen und bieten damit ein hohes Maß an Flexibilität und Präzision.

Neben der Wahl des Modells ist die Repräsentation der Zeitreihen ein kritischer Erfolgsfaktor. Je nach Quelle und Zielsetzung können Zeitreihen als rohe Zahlenreihen, als abstrahierte Ereignisfolgen in Prozessmodellen oder als statistisch verdichtete Merkmale vorliegen. In einem unserer Anwendungsfälle wurden Zeitreihen beispielsweise durch Mittelwerte, Varianzen und andere aggregierte Kennzahlen zusammengefasst, bevor sie einem Sprachmodell übergeben wurden. In anderen Szenarien müssen Zeitreihen zunächst aus unstrukturierten Texten oder multimodalen Quellen extrahiert und entsprechend kodiert werden.

Insgesamt zeigt sich: GenAI bietet ein breites Spektrum an Möglichkeiten, um Zeitreihendaten deutlich intelligenter zu nutzen. Sie erlaubt nicht nur eine tiefere semantische Erschließung der Daten, sondern auch eine flexible Anpassung an unterschiedliche Anwendungsfälle – von der schnellen Exploration bis zur hochspezialisierten Prognose.

Tabelle 3: Zeitreihen – Use Cases in der Übersicht

Branche/ Bereich	Use Case	Aufgabentyp	Datentyp & Datenstruktur	Modell oder Methode
Produktion & Industrie 4.0	❶ Prozessoptimierung für Effizienzsteigerung und Qualität	Prozessoptimierung für Effizienzsteigerung und Qualität	Diskretisierte Maschinen-Messdaten	Transformer mit Embedding, trainiert durch unüberwachtes Lernen
Prozessmanagement (branchenübergreifend)	❷ Eventverarbeitung für Effizienzsteigerung	Automatisierte Verarbeitung von Eventdaten	Sequenzielle Datenpunkte	Vortrainierte LLMs, Large Process Models
Energiewirtschaft	❸ Energieprognose für Verbrauchsoptimierung	Prognose von Energiedaten	Zeitreihen zu Erzeugung und Verbrauch von Energie	Überwachtes Lernen, vortrainierte Large-Timeseries-Modelle
Branchenübergreifend	❹ Agentengestützte interaktive Datenanalyse	Interaktive Visualisierung von Zeitreihendaten	Zeitreihendaten	LLMs mit RAG
Branchenübergreifend	❺ Positionsbestimmung für Standortoptimierung	Funkbasierte Lokalisierung	Funksignalzeitreihen	Selbst überwacht und vortrainierte Transformer Modelle
Produktion & Industrie 4.0	❻ Objekterkennung für Sicherheitsverbesserung	Radarbasierte Objekterkennung und Klassifikation	Range- (Entfernungs-) und Doppler-Daten	vortrainierte Large-Timeseries-Modelle und Nachtrainieren
Produktion & Industrie 4.0	❼ Maschinenüberwachung für Ausfallprävention	Sparsame Zustandsüberwachungen in industriellen Maschinenparks	Zeitreihen von maschinenintegrierter Sensorik	Vortrainiertes LLM und Nachtrainieren anhand aktueller Maschinendaten

2.3. Use Cases

2.3.1. Assistenzsystem zur Parameterwahl für Produktionsprozesse ❶

Die Modellierung komplexer Fertigungsprozesse erfordert komplexe Modelle. Ein in der Industrie häufig vorkommender Fertigungsprozess ist der Extrusionsprozess, eine Kombination komplexer chemischer und mechanischer Phänomene, der beispielsweise in der Kunststoffverarbeitung Anwendung findet. Die Wahl der richtigen Prozessparameter (z. B. Drehzahl, Temperatur von Heizelementen) ist erfolgskritisch und nicht trivial. Durch die Entwicklung eines Transformer-basierten Foundation-Modells, das auf Simulationsdaten trainiert und auf einzelne Maschinen angepasst wird, können wenige Maschinendaten sinnvoll zur Prozessoptimierung eingesetzt werden.

Physikalische Modelle bieten wertvolle Einblicke, sind für die Wahl passender Prozessparameter allerdings nur bedingt geeignet. Optimal wäre, wenn das Bedienpersonal gewünschte Eigenschaften des Produktionsprozesses, z. B. die Temperatur im Extruder, vorgeben könnte und ein Assistenzsystem passende Prozessparameter liefert. Physikalische Simulationen benötigen allerdings die Prozessparameter als Eingabe. Geeignete Parameter werden mithilfe dieser Modelle bisher iterativ und manuell ermittelt. Das bindet Fachpersonal und erschwert die Automatisierung, was den Einsatz von Demand-Side-Management und die kontinuierliche Anpassung der Parameter an den Machienenzustand verhindert.

Eine Studie des Fraunhofer Institut für Techno- und Wirtschaftsmathematik (ITWM) zeigt, dass datenbasierte Modelle diese Steuerungs-Aufgabenstellung meistern [15]. Aufgrund der Pluralität verschiedenster Materialien und Schneckengeometrien wurde der Ansatz mit Transformer-Modellen weiterentwickelt.

Basierend auf Simulationsdaten werden große Transformer-Modelle trainiert, die dann mit wenigen Messdaten gezielt an einzelne Maschinen angepasst werden (Fine-Tuning). Ergebnis ist ein Modell, das die Wirkmechanismen durch die Simulationsdaten erlernt hat. Die Ungenauigkeiten der Simulationen werden durch die Anpassung an Messdaten kompensiert.

Zum Trainieren werden Prozessparameter, Schneckengeometrien und gewünschte Temperatur- und Drucksequenzen in diskrete Werte eingeteilt. Das »Masking« erlaubt es, eine gewünschte Temperatur- und Drucksequenz vorzugeben,

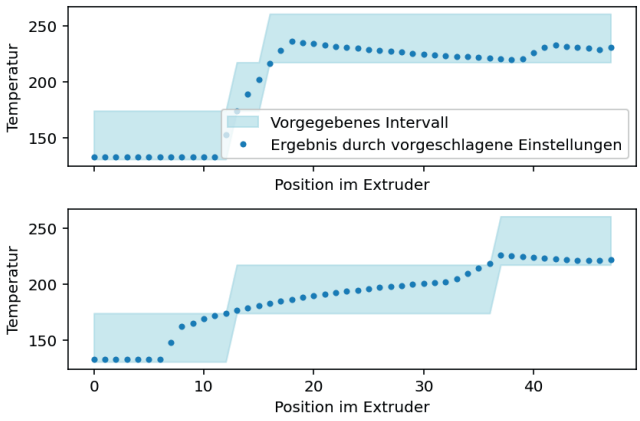


Abbildung 10: Die vorgegebenen Prozessparameter führen zum gewünschten Temperaturverlauf der Schmelze.
Quelle: Fraunhofer ITWM

sodass das angepasste Transformer-Modell die dazu passenden Prozessparameter ausgibt. Der Erfolg der Methode ist in Abbildung 10 gezeigt: Die vom Modell vorgeschlagenen Parameter führen zu Temperaturen im vorgegebenen Intervall. In der Anwendung entlastet ein hierauf basierendes Assistenzsystem Fachkräfte und reduziert den zeitlichen Aufwand sowie die Kosten, die durch Trial-and-Error oder Simulationen zur Ermittlung passender Prozessparameter entstehen.

Das ITWM hat die Grundlage eines Assistenzsystems generiert, um den komplexen Extrusionsprozess besser zu kontrollieren und den Zustand jeder einzelnen Maschine durch kontinuierliche Modellanpassung genau abzubilden.

So können schnell und ohne großen Aufwand wenige Maschinendaten sinnvoll zur Prozessoptimierung genutzt werden. Dieses Vorgehen lässt sich auch auf andere Produktionsprozesse anwenden.

Kontakt

Julia Burr
Fraunhofer ITWM
julia.burr@itwm.fraunhofer.de

2.3.2. Eventdaten als Treiber für moderne Process Intelligence ❷

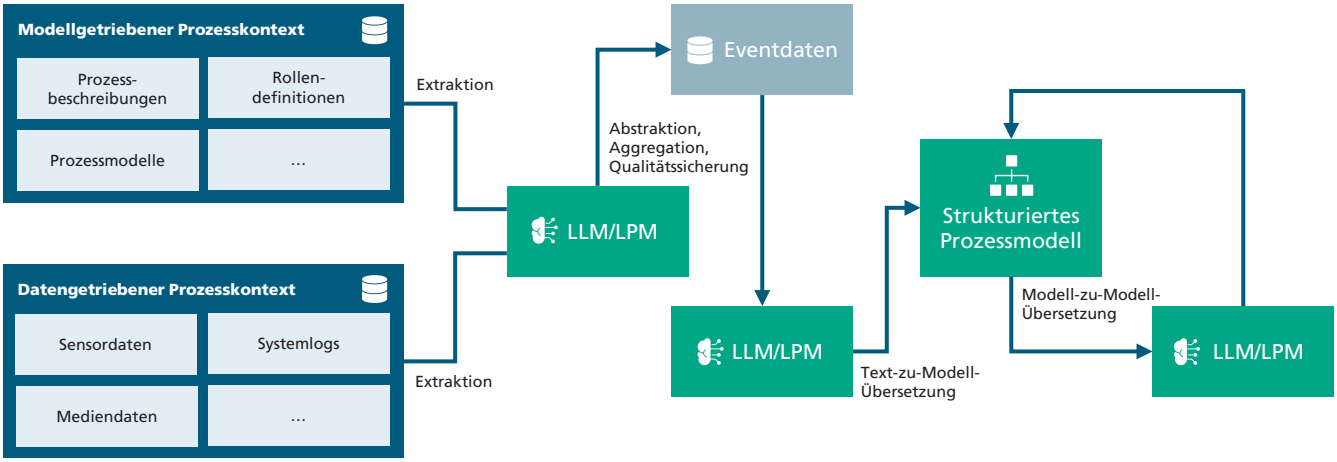


Abbildung 11: Referenzarchitektur von GenAI in Process Intelligence
Quelle: Fraunhofer FIT

Eventdaten sind das Fundament für die Realisierung von Process Intelligence, also die Analyse, Überwachung und Optimierung von Geschäftsprozessen. Mit ihren präzisen, zeitgestempelten Informationen liefern sie, was tatsächlich in unseren Geschäftsprozessen passiert – nicht nur, was laut Prozessmodell oder Betriebshandbuch passieren sollte. Sie bestehen aus sequenziellen Datenpunkten, die während der Prozessausführung gesammelt werden, und sind somit eine besondere Form von Zeitreihendaten.

Die Verarbeitung von Eventdaten stellt Unternehmen vor zentrale Herausforderungen. Diese Daten entstehen in großen Mengen und kommen aus unterschiedlichsten Systemen, was ihre Zusammenführung technisch anspruchsvoll macht. Dazu sind die Formate der Eventdaten oft uneinheitlich, sodass eine aufwendige Harmonisierung erforderlich ist, bevor sie sinnvoll analysiert werden können. Auch Datenqualität ist ein Thema – fehlende oder fehlerhafte Events können Analysen verfälschen.

Abbildung 11 veranschaulicht die Rolle von LLMs zur Verarbeitung und Analyse prozessbezogener Daten. LLMs reduzieren die Herausforderungen bei der Verarbeitung von Eventdaten deutlich, da sie unstrukturierte oder uneinheitliche Daten automatisch harmonisieren. Sie erkennen Muster, Anomalien und Zusammenhänge in großen Datenmengen schneller und präziser als herkömmliche Methoden und ermöglichen eine interaktive Verbesserung der Datenqualität und Analysegeschwindigkeit.

Neben der Verarbeitung von Eventdaten ist die Prozessmodellierung ein spannendes Einsatzfeld für LLMs. Hierbei werden LLMs z. B. zur automatisierten Übersetzung unterschiedlicher Prozessmodelltypen, wie Business Process Model and Notation (BPMN) und Unified Modeling Language (UML), eingesetzt. Ein weiterer Ansatz ist die Text-zu-Modell-Transformation, bei der Prozesswissen aus textuellen Beschreibungen extrahiert und automatisiert in strukturierte Prozessmodelle, wie z. B. BPMN, umgewandelt wird.

Domänenübergreifende Foundation-Modelle für Prozesse abstrahieren unterschiedliche Modelltypen und führen diese durch generalistische Repräsentationen des enthaltenen Prozesswissens zusammen. Multimodale Foundation-Modelle können zudem eventbezogene Informationen aus Texten, Bildern, Videos oder Audiodaten zu extrahieren und diese mit bestehenden Eventdaten zu korrelieren, um realitätsnähere Prozessmodelle zu generieren.

Im Rahmen des Forschungsprojekts »Anything-to-Log« wurden erste prototypische LLM-basierte Implementierungen zur Extraktion und automatisierten Abstraktion von Eventdaten aus unstrukturierten Video- und Textdaten entwickelt. Im Forschungsprojekt »QUAPRO« wurden LLMs eingesetzt, um die Qualität von Eventdaten automatisiert zu quantifizieren und zu verbessern. LLM-basierte Tools wie ProMoAI [16] und AIPA zeigen vielversprechende Ansätze zur automatisierten Erstellung und Verbesserung von Prozessmodellen. ProMoAI wandelt textuelle Fachbeschreibungen mithilfe spezialisierter GenAI-Modelle in strukturierte Prozessmodelle um. AIPA [17] verbessert die Verständlichkeit dieser Modelle durch eine interaktive, GenAI-gestützte Abfrageoberfläche.

Kontakt

Prof. Dr. Wolfgang Kratsch
Fraunhofer FIT
wolfgang.kratsch@fit.fraunhofer.de

2.3.3. Large-Timeseries-Modelle zur Prognose von Energiezeitreihen ❸

Die Vorhersage von Energieverbrauch und -erzeugung, wie zum Beispiel von Solar- und Windkraft oder wie viel Kundinnen und Kunden verbrauchen, ist entscheidend für die Planung und Beschaffung im Energiesektor. Diese Vorhersagen werden auch im Netzbetrieb immer wichtiger, da sie helfen, den Zustand des Stromnetzes zu überwachen. Bei Energievorhersagen für verschiedene Bereiche in großer Menge, wie zum Beispiel für Stadtteile mit erneuerbaren Energien und Elektrofahrzeugen, steigt der Aufwand für die Entwicklung und das Training spezifischer Prognosemodelle bzw. Fachmodelle erheblich. Methoden zum automatisierten Modellentwurf oder leistungsstarke Deep-Learning-Modelle können den Aufwand für Analyse und Engineering zwar prinzipiell reduzieren, benötigen jedoch weiterhin viele Trainingsdaten und damit verbundene Ressourcen.

Spezialisierte Foundation-Modelle für Zeitreihen oder LTMs (Large Timeseries Models) werden mit umfangreichen Zeitreihendaten, wie beispielsweise Google Trends, vortrainiert. Diese Modelle können Prognosen unter verschiedenen Bedingungen, wie unterschiedlichen Auflösungen oder Vorhersagezeiträumen, direkt liefern. Durch zusätzliches Nachtrainieren können sie noch höhere Genauigkeitsanforderungen erfüllen, abhängig von der Verfügbarkeit der Daten. Beispiele für solche spezialisierten LT-Modelle sind TimeGPT [18] (kommerziell), Lag-Llama [19] oder Timer [20] (beide Open-Source). Die Stärke von LTMs liegt in ihrer Fähigkeit, Vorhersagen ohne vorheriges Training, also als Zero-Shot zu treffen. So können sie auch bei schwierigen Datenlagen, wie fehlenden historischen Daten oder sensiblen Kundendaten, nützliche Prognosen liefern, ohne dass ein hoher zusätzlicher Aufwand erforderlich ist. Einige Studien [21] konnten erste vielversprechende Ergebnisse bei Lastprognosen aufzeigen; jedoch bestehen weiterhin Genauigkeitsverluste gegenüber Fachmodellen, besonders bei Mittel- und Langfristvorhersagen.

EMS-EDM Prophet [22] ist eine am Fraunhofer IOSB-AST entwickelte und seit über 25 Jahren im industriellen Einsatz befindliche Softwarelösung zur Ausführung von Energiemanagementaufgaben (z. B. Prognosen, Kraftwerkseinsatz, Bilanzausgleich) in Stadtwerken, Energieversorgern oder Netzbetreibern. Im EMS-EDM Prophet wurden erste prototypische Tests zur Day-Ahead-Vorhersage von Strompreisen mit einem Lag-Llama-Modell durchgeführt, das zwar vergleichbare Ergebnisse gegenüber einem klassischen Fachmodell

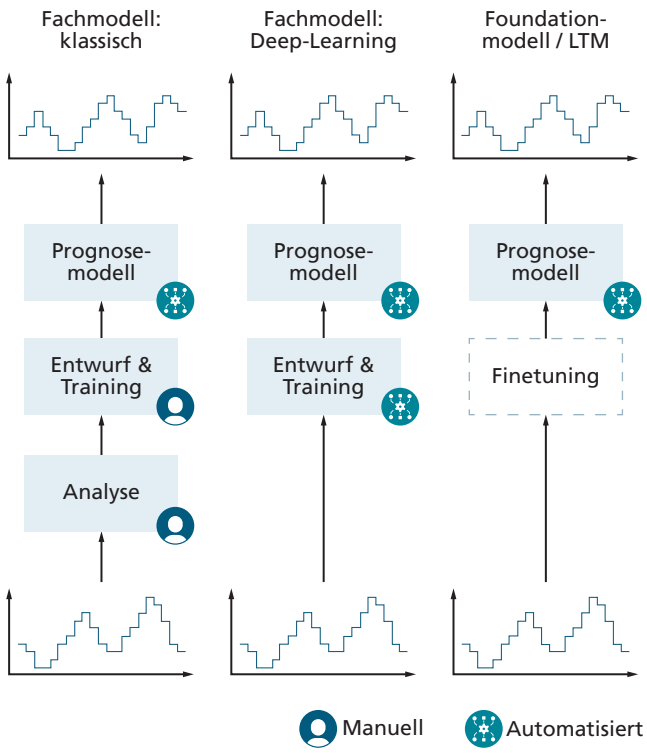


Abbildung 12: Vergleich Fachmodelle mit LTMs zur Vorhersage von Energiezeitreihen
Quelle: Fraunhofer IOSB-AST

aufweist, allerdings signifikant schlechtere Ergebnisse als ein Deep-Learning-basiertes Fachmodell erreicht. Die grundsätzliche Prognosefähigkeit von LTMs ist erkennbar, jedoch sollte der Einsatz von LTMs gegenüber Fachmodellen hinsichtlich der notwendigen Genauigkeit und dem akzeptablen Ressourcenaufwand sorgfältig abgewogen werden.

Weitere Tests von LTM-Modellen und deren Integration als Prognosedienst in EMS-EDM Prophet befinden sich aktuell in Arbeit. Auch ist die Einbindung von KI-Agenten zur Automatisierung von Zeitreihenanalysen ein in Zukunft vielversprechender Ansatz.

Kontakt

Dr. André Kummerow
Fraunhofer IOSB-AST
andre.kummerow@iosb-ast.fraunhofer.de

2.3.4. Agentengestützte interaktive Datenanalyse ❹

Die Zeitreihenanalyse spielt eine zentrale Rolle für das Verständnis messbarer Veränderungen im zeitlichen Verlauf. Sie ist in zahlreichen Anwendungsfeldern unverzichtbar, da sie zeitabhängige Dynamiken und zugrunde liegende Muster aufdeckt, die Grundlage für eine interaktive Datenanalyse und weitere strategische Entscheidungen sind. Obwohl bereits einige automatisierte Lösungen die Benutzerfreundlichkeit und Einfachheit der Zeitreihenanalyse erhöhen, erfordern diese Ansätze weiterhin ein gewisses Maß an Fachkenntnis.

Angesichts jüngster, bedeutender Fortschritte in der GenAI und der dadurch erfolgten Weiterentwicklung komplexer Sprachmodelle ergeben sich erweiterte Verwendungsmöglichkeiten für umfassende, jedoch leicht zugängliche Zeitreihenanalyse-Lösungen.

Der vom Fraunhofer FIT entwickelte Ansatz SAGE – Smart LLM Agents with GraphRAG for Time Series Exploration ermöglicht die Automatisierung von Zeitreihenanalysen in Kombination mit Domänen-, Fach- und Analysewissen, das mithilfe von Wissensgraphen und einer eigens entwickelten Ontologie namens SemTS [23] systematisch erfasst wird. Die Ontologie dient hierbei der semantischen Strukturierung von erfasstem Wissen bezüglich zu verarbeitender Zeitreihendaten. Beispiele für solches Wissen reichen von extrahierten Zeitreihenmerkmalen einschließlich struktureller Besonderheiten wie Anomalien oder Muster, bis hin zu domänenspezifischen Informationen wie Eventdaten oder Annotationen, die von Expertinnen und Experten bereitgestellt werden.

Durch die agentengestützte, automatisierte Ausführung von Datenanalysen und die Nutzung semantischer Informationen in Sprachmodellen ermöglicht SAGE die Formulierung von Anfragen in natürlicher Sprache, damit wertvolle Einsichten aus Zeitreihendaten gewonnen werden können (siehe Abbildung 13). Die zugrundeliegende Methodik nutzt dabei aufgabenspezifische Agenten, um optimierte Analysestrategien zu identifizieren, die Analysen unter Zuhilfenahme bereits zuvor abgeleiteten oder bereitgestellten Wissens ermöglichen. Neu gewonnene Einsichten werden zudem ebenfalls inkrementell

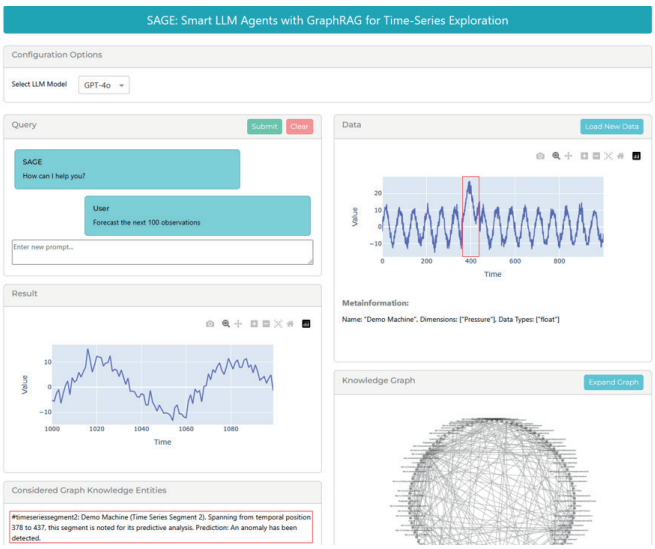


Abbildung 13: Interface der aktuellen Version des agentengestützen Datenanalysetools SAGE
Quelle: Fraunhofer FIT

in den Wissensgraph überführt, um weiterführende Analysen informativ anzureichern.

Unternehmen können hierbei ihre eigenen Zeitreihendaten nutzen, um klare und nützliche Erkenntnisse zu gewinnen, die die unmittelbare Entscheidungsfindung unterstützen. Dank der Möglichkeit, Datenanalysen über einfache Spracheingaben zu steuern, können sowohl Expertinnen als auch Anwender mit begrenzter Erfahrung in der Datenanalyse von dieser Funktionalität Gebrauch machen. Mit SAGE werden Unternehmen effizienter und wettbewerbsfähiger, da sie schneller auf Veränderungen im Markt reagieren und gezielte Entscheidungen treffen können.

Kontakt

Prof. Dr. Christian Beecks
Fraunhofer FIT
christian.beecks@fit.fraunhofer.de

2.3.5. Pretrained Transformer-Modelle zur funkbasierten Lokalisierung

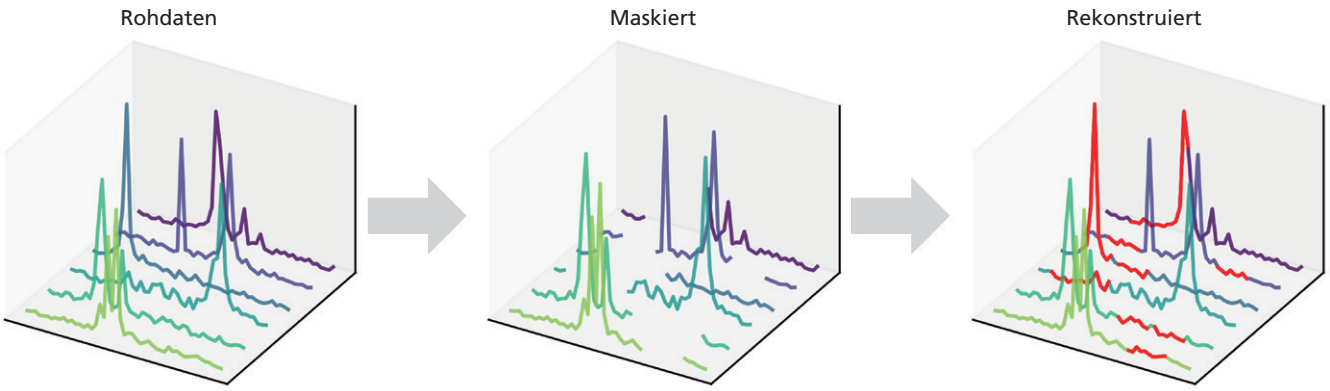


Abbildung 14: Unüberwachter Pre-Training-Prozess: Die Aufgabe des Modells ist es, vorher entfernte Signalteile zu rekonstruieren. Quelle: Fraunhofer IIS

Mit funkbasierter Lokalisierung können Objekte verortet und somit eine Vielzahl an Geschäftsmodellen ermöglicht werden. Während die Technologie im Außenbereich in Form von Global Navigation Satellite Systems (GNSS) etabliert ist, gewinnt sie, insbesondere in industriellen Innenraumumgebungen, wie z. B. Werkshallen, immer mehr an Bedeutung. Anwendungsfälle umfassen z. B. die Automatisierung und Überwachung von Produktionsprozessen in der Industrie oder die präzise Standortverfolgung von Patienten und Geräten im Gesundheitswesen.

Das Messprinzip basiert auf der Ausbreitung elektromagnetischer Wellen im Raum. Ein mobiles Tag sendet ein Funksignal, das an mehreren fest installierten Antennen im Raum empfangen wird. Die Signallaufzeiten werden geschätzt, und mittels geometrischer Berechnungen kann schlussendlich eine Position ermittelt werden. In der Praxis interagiert das Funksignal (vor allem in Innenräumen) jedoch mit der Umgebung – z. B. werden Signale an Wänden reflektiert. Infolgedessen erreichen die Signale die Antennen über mehrere Wege, sogenannte Mehrwegekomponenten (englisch: Multipath Components (MPCs)). Diese erscheinen dann wie zeitversetzte Kopien des auf dem direkten Weg empfangenen Signals. Kanalmessungen bilden die empfangene Signalstärke über der Zeit ab. Dabei manifestieren sich MPCs typischerweise als überlappende Signalspitzen.

Klassische Verfahren zur Funklokalisierung stoßen in komplexen Umgebungen mit starkem Mehrwegeempfang aufgrund starker Interferenzen oft an ihre Grenzen. MPCs liefern jedoch räumliche Information, die gezielt genutzt werden kann, um die Genauigkeit der Positionsbestimmung zu verbessern. Deep-Learning-Methoden eignen sich, um den vollen Informationsgehalt von Kanalmessungen auszuschöpfen, was jedoch oft mit erheblichen Kosten für die Datenakquisition einhergeht. Zwar lassen sich Funksignale in großem Umfang und geringem Aufwand erfassen, bisher eingesetzte Deep-Learning-Methoden sind jedoch auf große Mengen annotierter Daten angewiesen – d. h. zusätzlich zu den aufgezeichneten Funksignalen werden Referenzwerte (Labels) benötigt. Die Generierung dieser Labels ist meist mit erheblichem Aufwand verbunden, da sie hochpräzise Messungen mittels Referenzsystemen erfordert und somit kostspielig ist.

Ein vielversprechender Ansatz zur Überwindung dieses Problems sind Foundation-Modelle. Auf Funkdaten vortrainierte Foundation-Modelle können die Menge der benötigten annotierten Datenpunkte reduzieren und somit die Kosten senken. Das Fraunhofer Institut IIS konnte zeigen, dass ein nachtrainiertes Foundation-Modell die Menge der benötigten annotierten Datenpunkte auf ein Zehntel im Vergleich zu herkömmlichen Deep-Learning-Methoden reduzieren kann [24]. Abbildung 14 zeigt beispielhaft den unüberwachten Pre-Training-Prozess, der auf der Rekonstruktion fehlender Signalteile basiert: Das Modell lernt dabei zuerst, modifizierte Kanalmessungen zu rekonstruieren und entwickelt so ein Verständnis über die lokale Funkausbreitung. Ein anschließendes Nachtrainieren dieses Basismodells mit annotierten Trainingsdaten ermöglicht schließlich eine Positionsbestimmung.

Kontakt

Dr. Christopher Mutschler
Fraunhofer IIS
christopher.mutschler@iis.fraunhofer.de

2.3.6. Radarbasierte Objekterkennung und -klassifikation

In Mobilitätsanwendungen ist die Erkennung sich bewegender Objekte wie Fußgänger, Radfahrer, Fahrzeuge und unbemannte Luftfahrzeuge (UAVs) essenziell, da davon autonome Reaktionen abhängen (z. B. durch Fahrzeuge oder Roboter) (Abbildung 15). Entfernung und Geschwindigkeit werden über Range- und Doppler-Daten gemessen – basierend auf Laufzeit und Frequenzverschiebung reflektierter Radarsignale. Da Radar unabhängig von Licht, Wetter und Sichtverhältnissen arbeitet, ist es ideal für industrielle und kommunale Anwendungen geeignet.

Klassische Radarsysteme stoßen jedoch an Grenzen: Ein Fußgänger und ein Radfahrer mit ähnlicher Geschwindigkeit erscheinen gleich, UAVs werden durch ihre Größe und Geschwindigkeit oft falsch klassifiziert oder im Clutter übersehen. Umweltstörungen wie Reflexionen, Nebenkeulen-Artefakte und Rauschen beeinträchtigen zusätzlich die Datenqualität. Erste Lösungsansätze wurden mit Deep-Learning-Modellen verfolgt, die Radar-Rohdaten als ein- oder zweidimensionale Zeitreihen analysieren [25][26][27].

GenAI bietet neue Ansätze zur Datenaugmentation und zur direkten Verarbeitung von Radar-Zeitreihen durch Transformer-Modelle mit automatischer Zielerkennung. Das Fraunhofer IPA untersucht in einem aktuellen Projekt die Anwendung von GPT-ähnlichen Modellen, die mithilfe von LoRA effizient auf Standardhardware trainiert werden können [28]. Dabei wird der OpenRadar-Ground-Surveillance-Datensatz [29] verwendet, der Doppler-Spektren, Entfernung, Azimut und Geschwindigkeit aus Radarsensoren enthält [30]. Statt statischer Bilder werden die Radarsignale als Zeitreihen interpretiert, wodurch Modelle fein strukturierte Bewegungsmuster wie Mikro-Doppler-Signaturen lernen können – etwa zur Unterscheidung eines Fußgängers von einer Drohne oder einem Fahrzeug [31]. Diese Methode adressiert Artefakte und Rauschen, generalisiert auf neue Bedingungen und Zieltypen und eignet sich für Anwendungen in Produktion und Mobilität, etwa auf Robotern oder autonomen Fahrzeugen.

Diese Methode ist darauf ausgelegt, die genannten Artefakte und stark verrauschte Umgebungen zu bewältigen und dabei

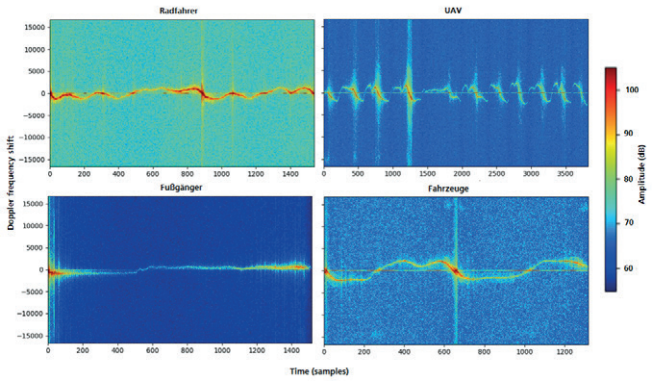


Abbildung 15: Doppler-Frequenzverschiebungen vs. Laufzeit für Radfahrer, UAV, Fußgänger, und Fahrzeuge. Erzeugt aus den Codebeispielen des OpenRadar-Datensatzes [29]. Quelle: Fraunhofer IPA

auf neue Bedingungen und Zieltypen zu generalisieren. Sie ist in der Produktion und in der Automobilindustrie einsetzbar – z. B. auf Robotern oder autonomen Fahrzeugen – überall dort, wo eine robuste, echtzeitfähige Identifikation erforderlich ist. Trotz dieser Vorteile muss die Antwortzeit der Methode für jeden konkreten Anwendungsfall sorgfältig berücksichtigt werden. Dieser Ansatz ermöglicht zudem Skalierbarkeit durch Few-Shot Learning, wodurch neue Objekte bereits mit wenigen Radarsignaturen erkannt und klassifiziert werden können. Zukünftige Erweiterungen beinhalten die Echtzeiterkennung von Anomalien, multimodale Sensorfusion (die Kombination von Daten aus verschiedenen Sensortypen (z. B. Radar und Kamera) sowie kontinuierliches Lernen zur Anpassung an neue Zieltypen und Umgebungen.

Kontakt

Alper Yaman, Ph.D.
Fraunhofer IPA
alper.yaman@ipa.fraunhofer.de

2.3.7. Sparsame Zustandsüberwachungen in industriellen Maschinenparks ❼

Produzierende Unternehmen besitzen oft große Maschinenparks mit gleichen oder ähnlichen Maschinen, bei denen fortlaufend Maschinen- und Prozessdaten erfasst werden. Dies ermöglicht eine Überwachung der Maschinen und ihrer Komponenten. Bei komplexen Maschinen erfolgt die Überwachung häufig KI-gestützt, insbesondere wenn hochfrequente Sensorik eingesetzt wird. Mögliche Zielsetzungen sind das frühzeitige Erkennen von defekten Komponenten, Verschleiß und Prozessanomalien.

Der Einsatz klassischer KI-Methoden zur Erkennung von Maschinen- und Prozessanomalien erfordert eine umfassende Menge gelabelter Sensordaten pro Maschinenkonfiguration. Während Gutdaten meist in ausreichender Menge vorliegen, sind Fehlerfälle oft unterrepräsentiert. Änderungen an Maschinenkomponenten können die Fehlererkennungsrate beeinflussen, weshalb die KI-gestützte Zustandsüberwachung neu trainiert werden muss.

Der Aufwand zur Inbetriebnahme KI-gestützter Systeme kann durch Transfer Learning oder Meta-Learning reduziert werden. Transfer Learning überträgt ein implementiertes und auf einem großen Datensatz vortrainiertes Modell auf die spezielle Aufgabe mittels weniger Daten. Meta-Learning nutzt spezielle KI-Modelle, wie Prototypical Networks oder Meta-SGD, um grundlegende Aufgaben, wie Klassifikation, über verschiedene Aufgaben mit jeweils wenigen Daten zu lernen. Das trainierte Meta-Modell wird dann auf die spezifische Aufgabe angewandt.

Trotz ihrer Eignung für Szenarien mit wenigen Daten benötigen diese Ansätze in der initialen Trainingsphase ausreichend viele gelabelte Daten der spezifischen Anwendung.

GenAI-Modelle nutzen umfassende Datensätze verschiedenster Bereiche beim initialen Training. Daher haben sie bereits eine breite Wissensbasis und können mit wenigen Datenpunkten auf neue Aufgabenstellungen eingelernt werden.

Besonders LLMs sind auf einer Vielzahl semantischer Daten eingelernt. Die Idee ist, Produktionsdaten in Sprache einzubetten, um sie als Eingabedaten für ein LLM zu nutzen. Das allgemeine LLM wird dann mit wenigen gelabelten Daten an die Domäne angepasst (»Feinabstimmung«).

Dieser Ansatz wurde erfolgreich zur Klassifikation von Kugellagerdefekten eingesetzt [32]. Dabei werden klassische Merkmalsberechnungen mit Prompting-Ansätzen kombiniert (siehe Abbildung 16), indem Merkmale aus dem Zeit- und Frequenzbereich extrahiert werden (z. B. Median, Varianz).

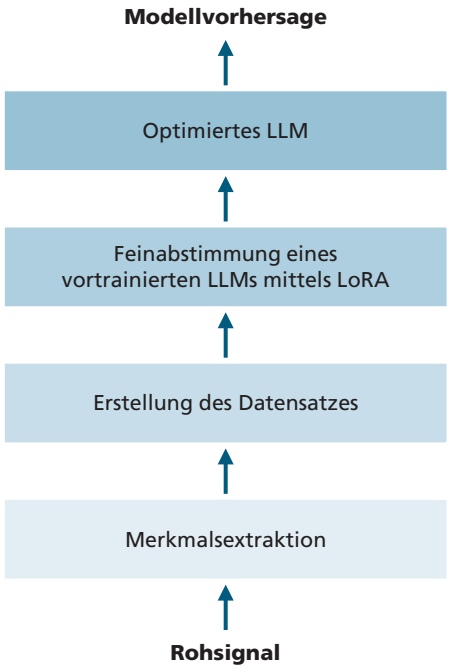


Abbildung 16: Konzept zum Einsatz von LLMs (nach Tao [32])
Quelle: Fraunhofer IPA

Diese Merkmale werden zur Feinabstimmung eines vortrainierten LLMs mittels Low-Rank Adaptation (LoRA) genutzt. Die extrahierten Merkmale werden dem LLM sowohl bei Training als auch im Produktivbetrieb in Textform übergeben. Bei der Inferenz prognostiziert das LLM den Lagerzustand und gibt ggf. eine Fehlerdiagnose aus.

Die Ergebnisse zeigen, dass dieser Ansatz auch bei komplexen Szenarien eine hohe Prognosegüte erreicht. Aufgrund der positiven Resultate der Lagerschadenserkennung ist angestrebt, den Ansatz auf andere Komponenten wie Pumpen zu übertragen. Pumpen weisen eine hohe Variantenvielfalt auf, was die Implementierung von Überwachungssystemen erschwert. Der MIMII-Datensatz [33] eignet sich zur Evaluation der Leistungsfähigkeit verschiedener Ansätze und erlaubt einen Vergleich zu Transfer- und Meta-Learning Ansätzen.

Kontakt

Christian Seidler
Fraunhofer IPA
christian.seidler@ipa.fraunhofer.de

3. Graphen für komplexe Netzwerke

3.1. Die Datenstruktur

Graphen sind effektive Hilfsmittel zur Darstellung komplex strukturierter Daten. Sie dienen dazu, die in Daten enthaltenen Beziehungen oder Netzwerke möglichst einfach zu beschreiben. Darauf aufbauend lassen sich in den Daten inhärente Muster, Trends und Zusammenhänge ableiten und analysieren. Mit Graphen lassen sich somit wertvolle Informationen in Daten aufdecken, die mit anderen Methoden möglicherweise nur unzureichend erkannt werden können.

Sowohl in der realen als auch in der digitalen Welt existieren diverse Beispiele für Graphen: das Verkehrsnetz als Menge von Verbindungsmöglichkeiten zwischen Orten (vgl. Abbildung 17), ein Marktplatz als Menge von Verbindungen zwischen Anbietern und Käuferinnen oder Moleküle als Menge von Verbindungen von Atomen. Auf diese Graphen lassen sich speziell entwickelte Analysemethoden anwenden, um beispielsweise vorherzusagen, welche Verkehrsmenge zu erwarten ist, welche Produkte ein Käufer als Nächstes kaufen könnte oder welche medizinische Wirkung ein bestimmtes Molekül besitzt.

Die Struktur eines Graphen setzt sich grundlegend aus zwei Komponenten zusammen. Diese werden als Knoten und Kanten bezeichnet. Jeder Knoten repräsentiert ein individuelles Objekt oder Element im Graphen. Die jeweiligen Beziehungen zwischen zwei Knoten werden als Kanten bezeichnet. Am Beispiel des Graphen der Verkehrsnetze stellen Straßensegmente die Kanten dar, während Kreuzungen die Knoten repräsentieren. Aufgrund der Menge an Knoten und der Art ihrer Verknüpfung über die Kanten können Graphen beliebig groß werden, was z. B. bei der Darstellung von Molekülen extreme Dimensionen annehmen kann.

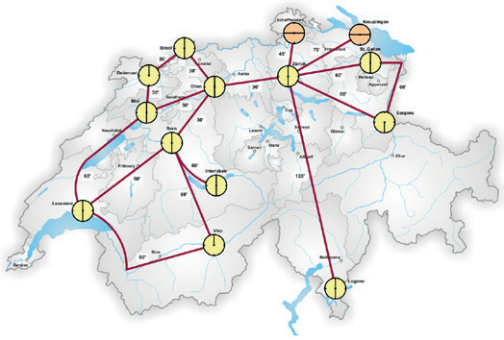


Abbildung 17: Ein Knotensystem dargestellt als Graph. Quelle: Wikipedia - Bernese_media

Die Beschreibung eines Graphen durch Knoten und Kanten kann um beliebige Merkmale ergänzt werden. In einem Verkehrsnetz sind dies z. B. die Anzahl der Abbiegespuren. Neben diesen numerischen Merkmalen können auch deskriptive Texte, wie zum Beispiel Produktbeschreibungen, als Knoten- oder Verbindungsmerkmale verwendet und in die Analyse einbezogen werden. Weiterhin wird zwischen gerichteten Graphen, bei denen die Verbindung eine bestimmte Richtung hat, und ungerichteten Graphen unterschieden. Für beide Typen kommen unterschiedliche Analysemethoden zum Einsatz.

Ein Beispiel zur Beschreibung gerichteter Graphen mit deskriptiven Texten ist das Resource Description Framework (RDF). Dieses beschreibt die Beziehung von Knoten und Kanten eines Graphen durch Aussagen der Form: Subjekt (Knoten) – Prädikat (gerichtete Kante) – Objekt (Knoten). Durch die Beschreibung eines Graphen im RDF-Format lassen sich logische Zusammenhänge darstellen, sodass man auch von einem Wissensgraphen spricht.

3.2. Warum GenAI für diese Datenstruktur?

Große Sprachmodelle (LLMs) haben die Verarbeitung natürlicher Sprache revolutioniert, indem sie komplexe Muster und Semantik in umfangreichen Textsammlungen effektiv erfassen können. Im Fokus steht dabei die Fähigkeit von LLMs, Zusammenhänge zwischen Tausenden unterschiedlicher Wörter besonders treffsicher zu repräsentieren. Eine analoge Aufgabe gibt es bei der Verarbeitung von Graphen. Daher besteht ein großes Interesse, auch die Techniken hinter großen Sprachmodellen auf Graphstrukturen zu übertragen.

Eine direkte Übertragung der Verfahren ist jedoch nicht möglich, da Graphen im Gegensatz zu Texten eine netzartige statt einer sequenziellen Struktur aufweisen. Texte werden in einer einzigen Richtung gelesen, während eine Straße sich an Kreuzungen verzweigt. Aufgrund dieser Besonderheiten wurden spezielle Ansätze (z. B. Graph Transformer und Graph Neural Networks) entwickelt, um die Knoten und Kanten von Graphen samt deren heterogenen Attributen und Nachbarschaftsbeziehungen zu verarbeiten.

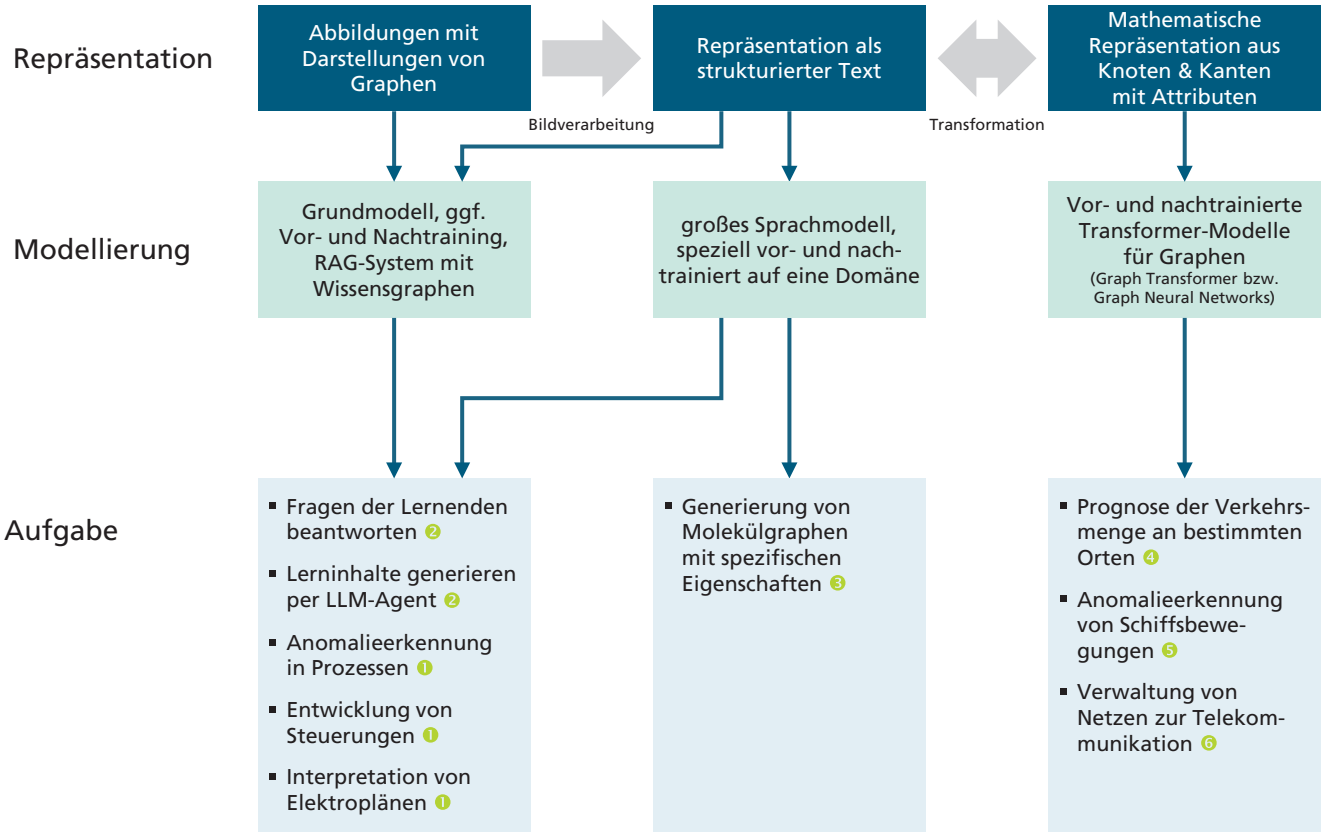


Abbildung 18: In der obersten Zeile werden die verschiedenen Repräsentationen der Graphen aufgeführt, die in unterschiedliche Typen von Embeddings transformiert werden. Während Texteinbettungen von Sprachmodellen verarbeitet werden, analysieren Graph Transformer die Embeddings der Knoten des Graphen. Dabei können externe Speicher wie RAG-Systeme und Wissensgraphen genutzt werden. Die unterste Zeile beschreibt die behandelten Use Cases.
Quelle: Fraunhofer IAIS

Neben diesen speziell für Graphen entwickelten Ansätzen kommen weitere Verfahren der GenAI im Umgang mit der Verarbeitung von Graphen zum Einsatz. Eine Übersicht aller Verfahren mit Bezug zu den hier präsentierten Use Cases zeigt Abbildung 18. Das jeweilige Verfahren ist abhängig von der Art der Repräsentation der Graphen:

- Graphen können sowohl in analogen als auch digitalen Dokumenten als Abbildung, z. B. als Elektroplan, repräsentiert werden. Für die weitere Verarbeitung müssen diese zunächst als Bild extrahiert werden. Aus solchen Bildern kann anschließend mithilfe großer multimodaler Sprachmodelle oder älterer Bilderkennungsmodelle eine textliche oder mathematische Repräsentation des Graphen extrahiert werden.
- Eine textliche Repräsentation der Knoten und Kanten eines Graphen kann durch spezielle genormte Beschreibungen erfolgen. Beispiele hierfür sind GraphML oder RDF auf XML-Basis. Der Vorteil solcher Darstellungen ist, dass sie direkt durch Sprachmodelle weiterverarbeitet werden können.
- Eine mathematische Repräsentation eines Graphen ist eine Nachbarschaftsmatrix, in der für jeden Knoten eine Zeile und Spalte existiert. Ein Eintrag zeigt an, ob eine Kante zwischen den Knoten vorhanden ist, und kann die Attribute der Verbindung enthalten.

Je nach Repräsentation kommen bei der Modellierung verschiedene Techniken zum Einsatz. Wenn die Netzwerkstruktur eines Graphen in Textform vorliegt, so kann dieser direkt mithilfe von Großen Sprachmodellen [34] analysiert werden. Dabei werden die Worte der Beschreibungssprache in Token überführt. Die numerischen Attributwerte der Knoten können entweder als Token oder unmittelbar als Zahlen kodiert werden. In beiden Fällen werden die LLMs auf einer Trainingsmenge von Graphen trainiert. Anschließend können sie fehlende Knotenmerkmale aus den Beobachtungswerten vorhersagen. Besonders relevant sind Graphen, deren Werte sich im zeitlichen Ablauf ändern können. In diesen Graphen können Sprachmodelle einerseits fehlende Werte einzelner Knoten vorhersagen, und andererseits zukünftige Werte prognostizieren.

Besonders vorteilhaft bei der Verwendung von Sprachmodellen ist, dass textuelle Merkmale der Knoten und Kanten direkt von einem entsprechend vortrainierten Sprachmodell berücksichtigt und mit der Graphenstruktur in Verbindung gesetzt werden können. Das bedeutet eine entscheidende Verbesserung der Analysemöglichkeiten im Vergleich zu rein numerischen Attributen. Es ist auch möglich, LLMs auf großen Graph-Trainingsdaten vorzutrainieren, und anschließend Prompts zu verwandten Graphen direkt zu beantworten [35]. Wie auch bei allgemeinen Sprachmodellen können Grapheninhalte in externe Datenbänke gespeichert und von LLMs genutzt werden (RAG). Ein Retrieval kann über eine Ähnlichkeitssuche stattfinden, die auch semantisch ähnliche Inhalte finden kann.

So wie Texte für Sprachmodelle in Teile (Token) zerlegt und für diese Vektoreinbettungen erzeugt werden, können auch für die Knoten in einem Graphen Vektoreinbettungen erzeugt werden. Dabei müssen die Position des Knotens im Graphen, seine Nachbarschaft sowie weitere relevante Merkmale und beschreibende Texte in die Knotenattribute integriert werden. Oft werden dabei mathematische Eigenschaften der Graphstruktur verwendet. In diesen sogenannten Graph-Transformern [36] reichert die Self-Attention die Knoteneinbettungen mit Informationen aus sämtlichen anderen Knoten und Verbindungen an, sodass in mehreren Ebenen des Netzwerks Embeddings entstehen, die die Knoteneigenschaften möglichst gut vorhersagen. Mithilfe von Sprachmodellen können durch Texte beschriebene Knotenattribute in Embeddings transformiert werden, die dann als Merkmale im Graph-Transformer berücksichtigt werden können. Zudem können Wissensgraphen als externe Wissensspeicher genutzt werden.

Der Graph-Transformer wird mit einer Trainingsmenge von Graphen trainiert, wobei z. B. maskierte Eigenschaften eines Knotens möglichst gut vorhergesagt werden sollen. Für die Auswertung wird ein neuer Graph eingegeben und es können neue Knotenwerte für Zielknoten prognostiziert und Knoten- oder Kantenmerkmale verändert werden. Graph-Transformer können aber auch fehlende Kanten oder Knoten prognostizieren und die Graphstruktur anpassen, um zum Beispiel zeitliche Veränderungen im Netzwerk zu berücksichtigen. Ihre Vorgänger, die Graph Neural Networks, beschränkten sich auf die lokale Aggregation von Merkmalen benachbarter Knoten.

Insgesamt bietet GenAI neue und leistungsstarke Architekturen für die Verarbeitung von netzartig strukturierten Unternehmensdaten. Die Auswahl eines geeigneten Ansatzes, der zu den vorhandenen Datenressourcen und der unternehmensspezifischen Aufgabenstellung passt, erfordert jedoch eine sorgfältige Evaluierung.

Tabelle 4: Graphen – Uses Cases in der Übersicht

Branche/ Bereich	Use Case	Aufgabentyp	Datentyp & Datenstruktur	Modell oder Methode
Produktion	❶ Interpretation von Elektroplänen und ähnlichen Diagrammen für die Anlagensteuerung und -überwachung	Engineering automatisieren (Schemata-Generierung) und bei Wartung und Instandhaltung assistieren (Schemata Interpretation)	Knoten: Komponenten wie Pumpen, Ventile, Sensoren Kanten: elektrische Leitungen, Luftkanäle, Rohre	Kombination von großen Sprachmodellen und RAG
Bildung	❷ Personalisierter Lernassistent im Bildungswesen	Lernenden Fragen beantworten, Lernfortschritte analysieren, Lerninhalte generieren	Graph mit RDF-basierten Daten zu Lerninhalten, -interaktionen und Lernenden	Transformer-Modelle wie Teuken, Mistral, zusätzlich RAG, Tool-Calling und Agenten
Pharma/Chemie	❸ Beschleunigung der Material- und Wirkstoffentwicklung in der Chemie- und Pharmaindustrie	Moleküle mit vorgegebenen Eigenschaften generieren	Graph mit Atomen und Atombindungen	Generative Transformer-Modelle
Mobilität	❹ Vorhersage von Verkehrsflüssen für die Mobilitätsplanung	Verkehrsmenge und Mobilität von Zielgruppen prognostizieren	Graph mit räumlichen und zeitlichen Sensormessdaten, Trajektorien, Verkehrsregeln etc. von Straßennetzen	Diffusionsmodelle und Graphtransformer: z. B. ImputeFormer
Sicherheit	❺ Schutz der kritischen Infrastruktur, Detektion illegaler Aktivität	Schiffspositionen und deren Aktivitäten überwachen	Raumzeitlicher Graph von Schiffspositionen	Graph Neural Networks
Telekommunikation	❻ Management von Telekommunikationsnetzen	Telekommunikationsnetze optimieren und verwalten	Repräsentation des Netzwerkgraphen durch RDF	Generative Transformer-Modelle

3.3. Use Cases

3.3.1. Sprachmodelle für die Anlagensteuerung und -überwachung ❶

Zunehmender Fachkräftemangel und der Bedarf an profitablen, skalierbaren Lösungen rückt GenAI auch im Bereich der Anlagensteuerung und -überwachung in den Fokus. Technische Zusammenhänge von Anlagen, wie sie beispielsweise in Elektroschaltplänen dargestellt sind, lassen sich als Graphen interpretieren und ermöglichen mithilfe von GenAI ein effizienteres Engineering und eine optimierte Produktion.

Während gängige Sprachmodelle wie GPT-4 die Generierung von Steuerungscode für Anlagen (insbesondere von structured text nach IEC 61131) schon recht gut beherrschen, fehlt eine umfassende Unterstützung für Programmiererinnen und Programmierern von speicherprogrammierbaren Steuerungen (SPS) auch bei komplexeren Aufgaben entlang des gesamten Softwareentwicklungszyklus.

Die Interpretation von Elektroplänen ist häufig mit hohen manuellen Aufwänden verbunden. Elektropläne können als Graphen interpretiert werden, wobei Knoten elektrische Komponenten und Kanten Strom- oder Signalleitungen darstellen (siehe Abbildung 19). Gleiches gilt für andere graphische Repräsentationen technischer Anlagen und Zusammenhänge, die als PDF, Bild, Text oder Schema vorliegen. Mit klassischen Bildverarbeitungs- und KI-Verfahren können die grafischen Abbildungen in ein textuelles Format, z. B. in eine Kantenliste, überführt werden. Über geeignete Einbettungen können diese Texte in RAG-Systemen kontextbezogen für LLMs abgerufen werden. Auf Basis der im RAG-System hinterlegten Beschreibungen über die Zusammenhänge der Anlage können SPS-Programmiererinnen und -Programmierer Hinweise für die Umsetzung der Steuerung oder Vorschläge für Unit Tests bekommen. Die Steuerungsentwicklung wird effizienter, weniger fehleranfällig und für Quereinsteiger zugänglicher, da Domänenwissen explizit abgefragt und automatisiert genutzt werden kann.

Ähnlich stellt sich die automatisierte Prozess- und Anlagenüberwachung dar, die grundsätzlich mit etablierten Verfahren der Anomalieerkennung beherrschbar ist. Auch hier fehlt eine umfassende Assistenz für Anlagenbedienende, die derzeit noch Handbücher mit schematischen Abbildungen von Fließschemata, Elektro- oder Pneumatikschemata studieren müssen. Zukünftig kann auf Basis einer erkannten Anomalie und den zugrunde liegenden Informationen aus Handbüchern, Schemata und der Wartungshistorie eine spezifische Handlungsempfehlung ausgegeben werden, die über die in RAG-Systemen hinterlegten Graphen die Verschaltung der

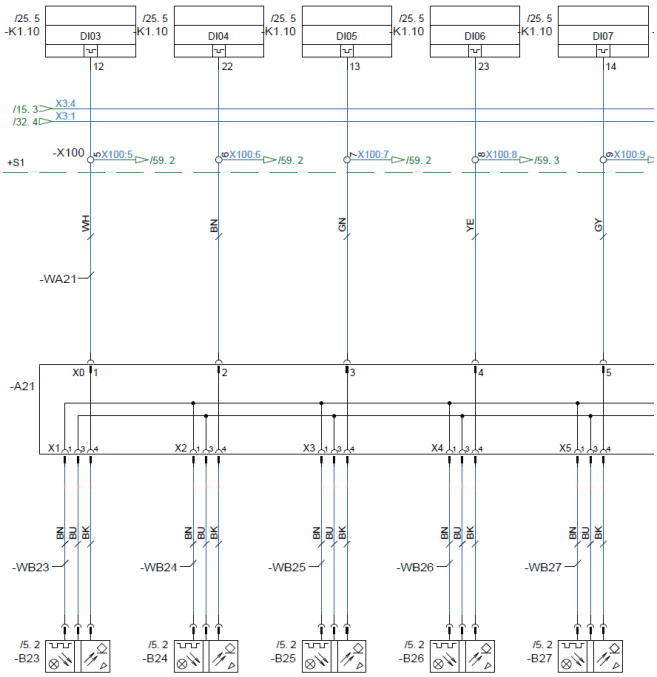


Abbildung 19: Ausschnitt aus einem Stromlaufplan
Quelle: Herbert Kannegiesser GmbH

technischen Komponenten und die Positionen der verbauten Sensorik berücksichtigt. Die verwendeten RAG-Systeme tragen in der Realisierung des Chunkings, der Auswahl der Embeddings und der Ähnlichkeitsbewertung den spezifischen Anforderungen an die Domäne und die Informationsgrundlage (z. B. Elektropläne) Rechnung. Die Fehlerbehebung und Wartung von Anlagen können somit effizienter, gezielter und durch weniger erfahrene Personen erfolgen.

Am Fraunhofer IOSB-INA werden für unterschiedliche Use Cases gemeinsam mit Industriepartnern Ansätze erarbeitet, um Informationen aus Schemata (Elektro- und Pneumatikschemata oder Fließbilder) für RAG-Systeme mit Sprachmodellen nutzbar zu machen.

Kontakt

Dr. Gesa Benndorf
Fraunhofer IOSB-INA
gesa.benndorf@iosb-ina.fraunhofer.de

3.3.2. Lernassistent für die Personalisierung in Lernplattformen

In der heutigen Bildungslandschaft stehen Unternehmen und Bildungseinrichtungen vor der Herausforderung, Lerninhalte effizient und personalisiert zu vermitteln. Lernmaterialien müssen hierzu oft individuell auf Lernende angepasst werden, um den unterschiedlichen Bedürfnissen und Lernstilen der meist heterogenen Zielgruppen gerecht zu werden. Nur so kann eine kontinuierliche Motivation und Selbstdisziplin gewährleistet werden, um den Lernfortschritt der Lernenden aufrechtzuerhalten.

Bei der Individualisierung von Lernmaterialien setzen LLM-basierte Lernassistenten an, die standardisiertes Wissen zu Lerninhalten und Lernenden durch große Sprachmodelle erschließen und bedarfsangepasst für die Nutzenden zur Verfügung stellen. Auf Basis des strukturierten Wissens in Lernplattformen und Bildungsökosystemen können verschiedene Anwendungsfälle durch den Einsatz spezialisierter LLM-Agenten realisiert werden.

Alle digitalen Lerninhalte sind dabei mittels RDF-Standards verknüpft und ermöglichen den Lernassistenten, persönliche Lernpfade flexibel und dynamisch an die individuellen Bedürfnisse der Lernenden anzupassen.

Ein weiterer Anwendungsfall ist die Beantwortung von Fragen der Lernenden per LLM-basierter Chatbots direkt in den Lernmanagementsystemen. Über ein RAG-System werden die persönlichen Lernstände der Nutzenden sowie die relevanten Lerninhalte inklusive der gesicherten Daten zur Urheberschaft und zu den einzelnen Verarbeitungsschritten aus verschiedenen Quellen, insbesondere aber aus der Lernplattform, abgerufen und für die Antwortgenerierung verarbeitet. Für personalisierte Empfehlungen können LLM-Agenten die auf RDF-basierenden Lerninteraktionen detailliert analysieren und Hinweise geben (siehe Abbildung 20).

Lernassistenten generieren auch neue Inhalte, wie personalisierte Übungsaufgaben, basierend auf individuellen Fortschritten und Vorlieben. Die Inhaltserstellung wird durch die Anreicherung der Metadaten durch LLM-basierte Analysen und beispielsweise die automatische Erkennung von vermittelten Kompetenzen und Schlagwörtern unterstützt.

Ein praktisches Beispiel ist »rAIbert«, ein LLM-basierter Lernassistent für das Kommando Heer und das Ausbildungskommando Heer der Bundeswehr aus dem Projekt KI in LMS [37] – einer Kooperation der Fraunhofer Institute FOKUS, FKIE und IOSB. rAIbert ist in bestehenden Lernmanagementsystemen omnipräsent verfügbar und nutzt große Sprachmodelle angereichert mit RAG für die Einbindung von Lerninhalten und KI-Agenten zur Analyse von Lernständen, um die Lernenden



Abbildung 20: Abbildung der persönlichen Lernstandsvorhersage Quelle: Fraunhofer FOKUS

individuell zu unterstützen. Dank der RDF-Beschreibungen kann der Chatbot detaillierte Interaktionsdaten verarbeiten, die Kursinhalte auswerten und nahtlos in der Plattform dargestellt werden. Der Einsatz von rAIbert als neuartiges ergänzendes GenAI-Werkzeug in der Lehre hat gezeigt, dass Lehrende viel Zeit bei der Bearbeitung von Standardaufgaben sparen und Lernende durch die personalisierte Ansprache und die adaptive Anpassung der Lerninhalte motivierter und effizienter lernen.

Kontakt

Dr. Christopher Krauss
Fraunhofer FOKUS
christopher.krauss@fokus.fraunhofer.de

3.3.3. De-Novo-Molekulardesign für gezielt ausgelegte Substanzen

Neue Materialien sind entscheidend für den Erfolg innovativer Technologien. Insbesondere im Kontext der Energiewende spielt die effiziente Suche nach neuen Substanzen eine zentrale Rolle. Häufig müssen Substanzen mit spezifischen Eigenschaften identifiziert werden. Organische Moleküle sind hierbei besonders relevant, da sie energetisch günstig produziert und oft einfach entsorgt werden können. Jedes Molekül kann dabei als Graph betrachtet werden, der aus einer Menge an Atomen gebildet wird.

Die Gesamtheit aller potenziell wirksamen Moleküle mit weniger als 30 Atomen, der sogenannte chemische Raum, wird auf 1060 geschätzt (zum Vergleich: Die Anzahl aller Atome im Universum liegt zwischen 1084 und 1089). Das heißt, dass wahrscheinlich für jede Anforderung ein passendes Molekül existiert. Gleichzeitig stellt die enorme Anzahl an Möglichkeiten jedoch eine sehr große Herausforderung dar, denn es können nicht einfach alle möglichen Moleküle auf deren Eigenschaften getestet werden. Daher birgt die gezielte Generierung von Kandidatenlisten mit geeigneten Molekülen ein großes Potenzial.

GenAI-Methoden bieten eine effektive Lösung für dieses Problem. Ähnlich wie bei großen Sprachmodellen wie ChatGPT wird eine Transformer-Architektur genutzt, um Moleküle in Form von Bindungsgraphen aus Atomen vorherzusagen, die den gewünschten Eigenschaften entsprechen. Dadurch kann der Suchprozess optimiert und die Anzahl der notwendigen Experimente im Molekulardesign erheblich reduziert werden.

Ein Beispiel für ein solches generatives Modell ist das im Rahmen des H2020-Projekts SONAR (2020–2023) entwickelte Sprachmodell LLaMol, das das De-Novo-Design von organischen Elektrolyten für Redox-Flow-Batterien unterstützt (siehe Abbildung 21) [38]. De-Novo-Design bezieht sich auf die computergestützte Entwicklung neuer chemischer Verbindungen oder Materialien, die noch nicht in der Natur vorkommen. LLaMol, basierend auf der Architektur des

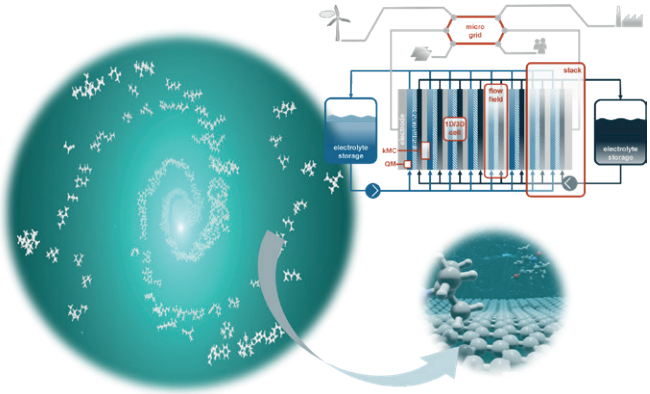


Abbildung 21: Schematische Darstellung einer Redox-Flow-Batterie einschließlich verschiedener Skalen und des chemischen Suchraums für Elektrolyte
Quelle: Fraunhofer SCAI

Sprachmodells LLama-2, unterstützt diesen Prozess, indem es hilft, gewünschte Eigenschaften oder Funktionen zu erreichen [39]. Es wurde auf 13 Millionen Molekülen aus öffentlichen Datenbanken trainiert. Als ersten Nachweis des Konzepts ermöglicht das Modell die Generierung von Molekülen unter Nebenbedingungen für vier kombinierbare beispielhafte Eigenschaften: den SAScore (Synthetic Accessibility Score), den logP-Wert (Maß für Lipophilie oder Hydrophobie), das Molgewicht und die Angabe eines Subfragments. Aktuell wird das Modell um weitere Eigenschaften erweitert und kann durch Nachtrainieren an spezielle Daten und De-Novo-Molekulardesignproblemstellungen von Kundinnen und Kunden angepasst werden.

Kontakt

Dr. Jan Hamaekers
Fraunhofer SCAI
jan.hamaekers@scai.fraunhofer.de

3.3.4. Prognose von Verkehrsmengen für die Standortplanung 4

Umfassendes Wissen über das Verkehrsgeschehen und das Mobilitätsverhalten der Bevölkerung ist für viele Branchen und Behörden von hoher Bedeutung. Von der Logistik über den Handel bis hin zur Infrastrukturplanung werden Daten über Verkehr und Mobilität umfassend für Planungszwecke eingesetzt.

Die empirische Erhebung dieser Daten ist jedoch sehr aufwendig. Oft ist die Erfassung lückenhaft und der Erwerb der Datensätze mit hohen Kosten verbunden. Für Planungszwecke sind diese Daten somit meist unzureichend. Deshalb werden seit vielen Jahren Modelle und flankierende Programme eingesetzt, die aufbauend auf den erfassten Daten vorhandene Lücken schließen und Prognosen über das zukünftige Mobilitäts- und Verkehrsgeschehen treffen. Diese Modelle sind jedoch teils aufwendig und unflexibel in der Erstellung bzw. Anwendung.

Zur Modellierung von Mobilität und Verkehr können neuerdings Graph-Transformer sowie Diffusionsmodelle verwendet werden (vgl. Vorgehen in rechter Spalte in Abbildung 18). Beide Verfahren können die Netzwerkstruktur des Verkehrsgraphen detailliert modellieren und bieten für diesen Anwendungsfall deutliche Vorteile gegenüber klassischen Verfahren wie Graph Neural Networks. Diffusionsmodelle generieren durch schrittweises Hinzufügen und Entfernen von Rauschen plausible Daten. Durch diesen stufenweisen Ansatz erzielen sie eine hohe Kontrolle über die Qualität und Vielfalt des Ergebnisses. Beide Verfahren sind im Ansatz komplementär und können für spezifische Fragestellungen kombiniert werden. Graph-Transformer können viele verschiedene Daten wie Verkehrssensoren, GPS-Daten oder Wetterdaten verarbeiten und sind daher ideal für die Vorhersage von Verkehrszuständen geeignet. Diffusionsmodelle können dagegen speziell für die Ausbreitung von Verkehr eingesetzt werden.

Bei beiden Ansätzen wird das Straßennetz als gerichteter Graph repräsentiert und Merkmale wie Abbiegeinformationen oder relevante Orte von Interesse wie Bahnhöfe oder Einkaufszentren hinzugefügt. Empirische Zählungen (sogenannte Frequenzen) durch Verkehrssensoren oder GPS-Tracks von Fahrzeugen und Personen werden Orten (Knoten/Kanten)

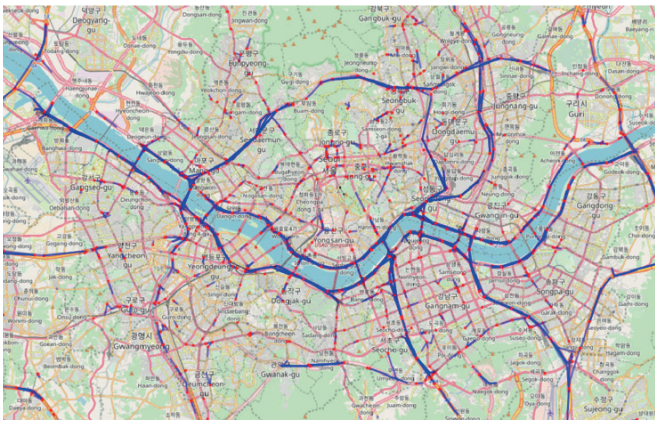


Abbildung 22: Ein Graph kann das Straßennetz einer Großstadt beschreiben.
Quelle: ResearchGate [40]

zugeordnet (vgl. Abbildung 22 [40]). Im konkreten Anwendungsfall soll die durchschnittliche Anzahl von Personen vorhergesagt werden, die beliebige Orte zu einer bestimmten Zeit passieren.

Die lückenhafte Datenbasis muss also auf das Straßennetz des Anwendungsgebiets hochgerechnet werden. Dabei werden die Struktur des Straßennetzes, mögliche Verkehrsquellen und sozioökonomische Daten der Bevölkerung einbezogen. Da sich Graph-Transformer und Diffusionsmodelle bei der Prognose von Verkehrsströmen bewährt haben, evaluiert das Fraunhofer IAIS spezielle Varianten zur Darstellung und Ergänzung fehlender Daten und berücksichtigt Methoden zur Analyse sehr großer Graphen. Die Anwendung dieser Modelle verspricht eine wesentlich bessere Erfassung der Verbindungsstruktur bei der Hochrechnung von Verkehrsflüssen und eine erhebliche Beschleunigung der gesamten Frequenzprognose.

Kontakt

Daniel Schulz
Fraunhofer IAIS
daniel.schulz@iais.fraunhofer.de

3.3.5. Eine flexible Datenstruktur für effiziente Schiffsüberwachung 5

In der maritimen Überwachung wird der Schiffsverkehr auf verdächtige Bewegungsmuster überwacht.

Aufgrund des hohen Verkehrsaufkommens auf den Ozeanen ist das für Menschen eine große Herausforderung. Zur Unterstützung werden GenAI-Methoden eingesetzt, die die Aufmerksamkeit auf auffällige Situationen lenken, beispielsweise illegale Fischerei, Sabotage oder Gefährdungen für den Seeverkehr. Sie werden auf Basis von Schiffsbewegungsdaten und zusätzlichen Informationen, wie Wetterdaten oder Standorten von Unterseekabeln oder -pipelines, modelliert.

Da es viele verschiedene Datentypen aus vielfältigen Anwendungen gibt, ist die Entwicklung von Modellen sehr komplex. Die bisherigen Ansätze sind auf bestimmte Anwendungen zugeschnitten und schwer auf neue Gegebenheiten und Anwendungen erweiterbar. Ein vielversprechender neuer Ansatz besteht darin, eine flexible Datenstruktur zu verwenden, um eine breite Palette von Situationen abzubilden und diese mit geringem Aufwand an neue Gegebenheiten anzupassen.

Ein solches Konzept wurde vom Fraunhofer IOSB entwickelt, das Zeitreihen durch gerichtete Graphen modelliert. Das Automatic Identification System als Quelle für Schiffsbewegungsdaten überträgt regelmäßig kinematische Informationen wie Position, Geschwindigkeit und Kurs sowie statische Daten wie Schiffstyp und Endziel. Diese Informationen helfen bei der Kollisionsvermeidung und der Überwachung des Schiffsverkehrs. Die Daten liegen in Form einer Liste von Meldungen vor und können hinsichtlich der kinematischen Informationen als Zeitreihen verstanden werden. Durch ihre Verknüpfung zu einem Graph können Muster und auffällige Situationen erkannt werden.

Für die Erkennung werden zunächst zeitlich geordnete Ketten von Knoten aus den Positionsmeldungen zu je einem Schiff in einer klassischen Graphstruktur gebildet (vgl. Abbildung 23). Die Interaktion zwischen den Schiffen wird durch Verbindungen zwischen Knoten ihrer Ketten, die in räumlicher und zeitlicher Nähe zueinanderstehen, repräsentiert. Der resultierende

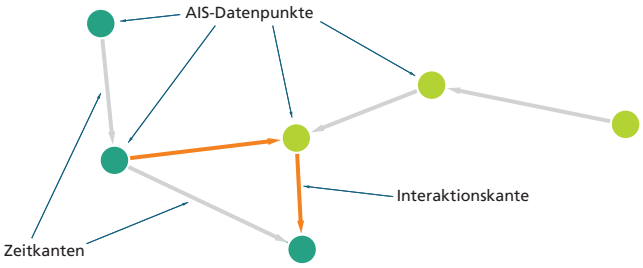


Abbildung 23: Räumlich-zeitlicher Graph für Schiffsbewegungen
Quelle: Fraunhofer IOSB

Graph kann nun durch Graph Neural Networks verarbeitet werden, um interessante Situationen zu detektieren.

Der hier vorgestellte Ansatz stellt ein neuartiges und flexibles methodisches Framework zur Betrachtung räumlich-zeitlicher Daten dar. Es ist zum Beispiel unabhängig von der Anzahl der Schiffe. Zudem lässt sich das Framework leicht und damit kosteneffizient auf neue Anwendungsfälle anpassen. Letztlich kann das Framework als Grundlage für die Entwicklung weiterer Anwendungen mit fortgeschrittenen Modellen wie Graph-Transformern oder Sprachmodellen dienen.

Die Strukturierung von Zeitreihendaten als Graphen und die Auswahl geeigneter KI-Modellarchitekturen wird zurzeit im Rahmen des EU-Projekts VIGIMARE zum Schutz kritischer Infrastruktur von Fraunhofer IOSB entwickelt und erprobt. Perspektivisch soll der Ansatz auf andere Felder wie Produktionsüberwachung übertragen werden. Auch hier werden Daten aus unterschiedlichen Quellen und verschiedenen Typs gemeinsam verarbeitet.

Kontakt

Matthis Leicht
Fraunhofer IOSB
matthis.leicht@iosb.fraunhofer.de

3.3.6. Flexibles Management für Telekommunikationsnetzwerke 6

Die Entwicklung der Telekommunikationsnetze hat sich seit ihren Anfängen dramatisch beschleunigt. Die Netzwerke der nächsten Generation, also der 5G- und 6G-Netze und ihrer Nachfolger, werden durch die Einbindung vielfältiger und heterogener Technologien komplexer. Die Einführung nicht öffentlicher Mobilfunknetze (NPN) ermöglicht es Organisationen, ihre eigenen Netzwerke zu betreiben, um bestimmte Anwendungsfälle zu realisieren. Aufgrund der zunehmenden Entflechtung des Netzes werden Hardwarefunktionen in Softwarekomponenten verlagert, die auf handelsüblicher Hardware ausgeführt werden können.

Die Verwaltung und Optimierung derart komplexer Systeme sind aufwendig, zeitintensiv und erfordern nicht nur von Expertinnen und Experten große Fachexpertise. Traditionell werden hierfür richtlinienbasierte Verwaltungs- und Kontrollsysteme (PBMC) genutzt, um die Komplexität zu verringern. Solche Ansätze erweisen sich jedoch als kostspielig, da eine Vielzahl von Komponenten getestet werden müssen, um den optimalen Betriebszustand des Systems zu erreichen.

Künstliche Intelligenz ist in der Lage, die Netzverwaltung zu vereinfachen. Zu diesem Zweck enthält die Literatur spezifische Modelle, die an bestimmte Datenströme zur Lösung eines Problems gebunden sind. Aufgrund der Vielfalt im Ökosystem der Telekommunikationsnetze müssen mehrere solcher Modelle erstellt werden und miteinander interagieren, was zu einer Verlagerung der Komplexität vom PBMC zum Modellmanagement führt. Für das Management hat die Anwendung der Transformer-Architektur in letzter Zeit zu verbesserten Lösungen geführt.

Um Netzwerkdaten in natürlicher Sprache darzustellen, braucht man einen klaren Ansatz. Wissensgraphen sind eine Art, Daten darzustellen, die sowohl für Menschen als auch für Maschinen verständlich sind. Standardisierte Ontologien helfen, Wissen eindeutig darzustellen. Graph Neural Networks (GNN) können diese Darstellung gut lernen. Neue Modelle, die auf der Transformer-Architektur basieren, bieten die Fähigkeit, aus Wissensgraphen fundierte Schlussfolgerungen zu ziehen.

Fraunhofer FOKUS entwickelt hierzu ein spezialisiertes Transformer-Modell KG-GPT (siehe Abbildung 24), das

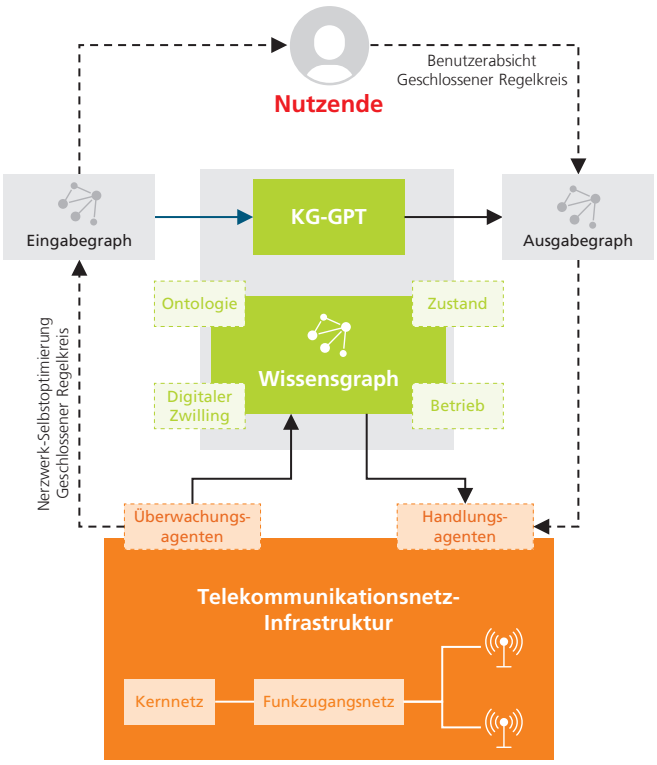


Abbildung 24: KG-GPT Funktionsdiagramm [42]
Quelle: Fraunhofer FOKUS

Wissensgraphen für Informationen über den Systemzustand nutzt. Der Wissensgraph wurde in einem Standardformat ausgedrückt, sodass die Netzwerkelemente asynchron bearbeitet werden können. Parallel dazu können die Transformer-Modelle auf den Wissensgraph zugreifen, wodurch die Graphen zum einheitlichen Kommunikationsmedium zwischen verschiedenen Modellen im Netzwerkmanagement werden.

Kontakt

Varun Gowtham
Fraunhofer FOKUS
varun.gowtham@fokus.fraunhofer.de

4. Digitale Zwillinge

4.1. Die Datenstruktur

Was sind Digitale Zwillinge?
Digitale Zwillinge sind virtuelle Repräsentationen physischer Systeme, die es ermöglichen, deren Verhalten und Zustand abzubilden. Neben Simulationen nutzen Digitale Zwillinge dazu ganz unterschiedliche Arten von Daten und Informationen, um beispielsweise Muster, besondere Ereignisse oder Anomalien zu erkennen. Hieraus lassen sich dann Schlussfolgerungen für das reale System ziehen ohne den tatsächlichen Prozess beeinflussen zu müssen. Digitale Zwillinge können für verschiedenste Anwendungsfälle entwickeln werden, wobei in diesem Kapitel konkret auf die in Abbildung 25 dargestellte Formen eingegangen wird.

Für alle Anwendungsfälle gilt, dass digitale Zwillinge die Möglichkeit bieten, physische Systeme in verschiedenen Detaillierungsgraden zu beschreiben: Wird früh in der Projektphase mit der Entwicklung eines Digitalen Zwillings begonnen, können schon in der Design- und Herstellungsphase Optimierungen am realen System vorgenommen werden. Im Betrieb selbst können Digitale Zwillinge auf aktuelle Prozesszustände zugreifen und durch kontinuierliche Datenauswertung proaktiv in das reale System ergreifen. Hierdurch lassen sich unter anderem die die Prozesseffizienz steigern, die Sicherheit des Systems verbessern oder auch deren Lebensdauer verlängern. Am Ende des Lebenszyklus des physischen Systems können digitale Zwillinge dann noch wichtige Informationen für das Recycling oder die Entsorgung bereitstellen. Dies ermöglicht eine ressourcenschonende Handhabung und fördert eine zirkuläre Wirtschaft, in der Materialien effizient wiederverwendet werden können.

Maschinen: In der industriellen Fertigung werden Digitale Zwillinge zur Überwachung und Steuerung von Maschinen eingesetzt, indem sie Daten über Design, Betrieb und Zustand sammeln.

Ein wichtiger Aspekt der Digitalen Zwillinge im Kontext Industrie 4.0 ist die Interoperabilität, die es ermöglicht, verschiedene Maschinen und Systeme miteinander zu vernetzen. Dadurch können Daten in Echtzeit ausgetauscht werden, was die Effizienz steigert und eine bessere Entscheidungsfindung im gesamten Lebenszyklus unterstützt. Beispielsweise erkennen sie Anomalien wie unerwartete Vibrationen oder Leistungsabfälle und helfen, proaktive Wartungsstrategien zu entwickeln, um Ausfallzeiten zu minimieren.

Mithilfe der Digitalen Zwillinge können Unternehmen fundierte Entscheidungen treffen, um Kosten zu minimieren und eine nachhaltige Ressourcennutzung zu gewährleisten.

Produkte: Digitale Zwillinge von Produkten ermöglichen es, deren gesamten Lebenszyklus zu verfolgen, vom initialen Design über die Herstellung und den Einsatz bis hin zur Entsorgung. Diese digitalen Modelle erfassen Daten über die Nutzung, den Zustand und die Leistung der Produkte. Hersteller können diese Informationen nutzen, um die Produktqualität zu verbessern, Rückrufaktionen zu vermeiden und personalisierte Dienstleistungen anzubieten. Ein Beispiel ist ein Digitaler Zwilling eines Fahrzeugs, der Daten über Fahrverhalten, Wartungsbedarf und potenzielle Probleme sammelt und analysiert. Diese Informationen können dazu beitragen, die Sicherheit und Zuverlässigkeit des Fahrzeugs zu erhöhen.

Gebäude: Ein Digitaler Zwilling eines Gebäudes kann verschiedene Aspekte abbilden und umfasst beispielsweise die Architektur, die physische Struktur und die technischen Systeme wie Heizung oder Lüftungsanlage. In der Praxis wird der Digitale Gebäudezwilling hauptsächlich dazu genutzt, die Interoperabilität zwischen verschiedenen Nutzenden während der Konstruktion zu gewährleisten; Architekten, Planerinnen und Baufirmen können sich so besser koordinieren. In einer hohen Entwicklungsstufe kann der Digitale Zwilling zudem zu Simulationen des Energieverbrauchs, der Luftqualität oder dem Verhalten von Nutzern im Gebäude eingesetzt werden. Zusammengefasst helfen Digitale Zwillinge sowohl bei der Konstruktion von Gebäuden als auch nach deren Erstellung indem sie z. B. Gebäudeverwaltungen helfen, Betriebsstrategien zu entwickeln und Wartungsarbeiten zu planen.

Menschen: Neben Digitalen Zwillingen von technischen Systemen können Digitale Zwillinge von Menschen erstellt werden, um menschliches Verhalten abzubilden und vorherzusagen. Diese Digitalen Zwillinge sind insbesondere für die Mensch-Maschine Interaktion wertvoll, um die Benutzungsschnittstellen besser an die Anforderungen und Intentionen der Nutzenden anzupassen. Bei sicherheitsrelevanten Bedienschritten (Steuern von Maschinen in Industrie-5.0-Anwendungen und von Fahrzeugen mit Driver-Monitoring-Systemen)

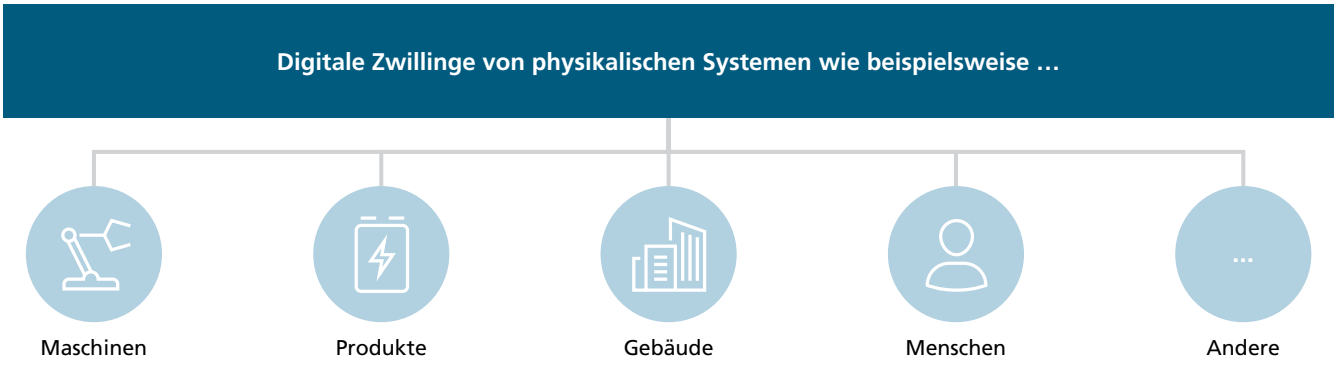


Abbildung 25: Ausprägungsformen von Digitalen Zwillingen
Quelle: Fraunhofer IAIS

können Eingabefehler vorhergesagt und verhindert werden. Für den Digitalen Zwilling eines Menschen sind Daten über vergangenes und aktuelles menschliches Verhalten notwendig. Je nach Anwendungsziel sind unterschiedliche Daten nötig. Viele Anwendungen profitieren von einem Modell des Menschen aus komplexen Datenstrukturen. Diese beschreiben den Menschen (Körperpose, Augenöffnung, physiologische Daten, Sprache, etc.), sein Verhalten und Interagieren (Lenkwinkel, Eingaben, Bedienschritte, etc.) sowie Kontextinformationen (Geschwindigkeit, Automationslevel, Wetter etc.). GenAI ist die erste Technologie, die diese unstrukturierten Datenklassen und Datenmengen in einem holistischen System fusionieren kann: einem universellen Digitalen Zwilling des Menschen.

4.2. Warum GenAI für diese Datenstruktur?

GenAI bietet eine herausragende Lösung für komplexe Datenstrukturen, die im Kontext Digitaler Zwillinge genutzt werden, da diese besonders dazu geeignet ist Strukturen und Komponenten von Systemen aus verschiedenen Datenquellen zu erkennen und zu extrahieren. Beispielsweise können sogenannte Transformer-Modelle, die die natürliche Sprachverarbeitung unterstützen, genutzt werden, um den manuellen Arbeitsaufwand bei der Datenvorbereitung, Informationsgewinnung und Strukturierung erheblich zu reduzieren.

Dies ermöglicht es, unstrukturierte Daten auf vordefinierte Strukturen in standardisierten Digitalen Zwillingen abzubilden. Zum Beispiel wird die Verwaltungsschale (VWS) (Asset Administration Shells; AAS) im Bereich der Industrie 4.0 eingesetzt. Die Expertise, die zur Erstellung und Nutzung von Digitalen Zwillingen erforderlich ist, wird durch die natürliche Sprache signifikant reduziert. Eine weitere zentrale Datenstruktur für komplexe Systeme bzw. Anlagen sind Wissensgraphen, mit deren Hilfe auf selbsterklärende Weise die Strukturen und Funktionen von Systemen beschrieben werden können. Wissensgraphen können ebenfalls auf Basis von vorhandenen Systembeschreibungen von GenAI erzeugt werden und dienen als zentrales Fundament für die Erstellung von Digitalen Zwillingen.

Einer der Hauptvorzüge von Transformer-Modellen ist zudem ihre Fähigkeit zur Graph-Strukturanalyse, sodass die erfassten Daten in verschiedenen Kontexten verwendet werden können, etwa für Plausibilitätsprüfungen der automatisch erzeugten Wissensgraphen oder Konformitätskontrollen mit Normen und Validierungen. Ein häufig auftretendes Problem bei der Erzeugung von Modellen ist die Qualität aus den Eingangsdaten: Unvollständigkeit, fehlerhafte oder widersprüchliche Daten treten oft auf und stellen formale Verfahren vor unlösbare Probleme. Hier haben Transformer-Modelle deutliche Vorteile.

Transformer-Modelle helfen aber nicht nur dabei, Digitale Zwillinge zu erstellen und zu validieren, sondern ermöglichen es auch intuitiv, mit diesen in einer dialogartigen Benutzerführung zu interagieren. Nutzende müssen sich nicht mehr mit komplexen Abfragen (z. B. über APIs oder SPARQL) auseinandersetzen; können die natürliche Sprache verwenden, die von einem Transformer-Modell in die jeweilig benötigte Abfragesprache übersetzt wird. Zudem können Transformer-Modelle bei der Konfiguration des automatischen Echtzeitzugriffs auf Betriebsdaten helfen, was die Einrichtung und Pflege der Digitalen Zwillinge deutlich verbessert.

Tabelle 5: Digitaler Zwilling – Uses Cases in der Übersicht

Branche/ Bereich	Use Case	Aufgabentyp	Datentyp & Datenstruktur	Modell oder Methode
Immobilien	❶ Digitale Zwillinge für effiziente Gebäude-bewirtschaftung	Wissensgraphen als Grundlage für effiziente Gebäudebewirtschaftung	Gebäudedokumentation in Form von Text, Bilder, Tabellen; RDF-Graphen	Transformer, LLM-RAG
Automotive	❷ Digitaler Zwilling von Fahrzeuginsassen	Abbild Insassen und Gesamtinnenraumsituation durch Videodaten und weiteren Datenquellen	Optische Sensorik; Sensorik zur Zustands-erfassung, Kontext der Fahrzeuginsassen Dokumente, Softwaresysteme	Computer Vision, LLM, Multimodale Modelle
Diskrete Fertigung und Prozessindustrie	❸ Der digitale Produktpass	Erstellung und Pflege von Verwaltungsschalen generieren	OPC UA-Datenmodelle	Transformer, Kontextuelles Verstehen Textgenerierung,
Industrielle Maschinen	❹ Semantisches Daten-mapping bei Digitalen Zwillingen	Informationsextraktion und Zuordnung (Mapping)	Asset Administration Shell; Metadaten	Transformer, LLM-RAG

4.3. Use Cases

4.3.1. Digitale Zwillinge für effiziente Gebäudebewirtschaftung ❶

Unternehmen in Deutschland stehen vor der Aufgabe, bis zum Jahr 2045 Klimaneutralität zu erreichen. Dabei spielt der Gebäudesektor eine zentrale Rolle, denn insbesondere Bestandsgebäude tragen erheblich zum Energieverbrauch und zu CO_2-Emissionen bei. Die Steuerung von Heizung, Lüftung und Klimaanlagen ist dabei ein relevantes Thema, falsche Einstellungen und kleine Havarien können schon 10 bis 30 Prozent Energieverlust verursachen.

Die Herausforderung ist, diese Anlagen so zu optimieren, dass sie netzdienlich und vor allem effizient im Umgang mit volatilen erneuerbaren Energiequellen sind. Dies erfordert eine sorgfältige Planung der Gebäude, oder bei Bestandsimmobilien eine nachträgliche Ertüchtigung, um Funktionen wie die Optimierung von Betriebsfahrplänen zu ergänzen. So ergibt sich ein erheblicher Automatisierungsbedarf, der zudem durch die Vielzahl an herstellerspezifischen Gebäudeautomationssystemen, Datenprotokollen und -schemata erschwert wird.

Ein Lösungsansatz, um die Interoperabilität der benötigten Automatisierungssysteme sicherzustellen, ist der Einsatz standardisierter Digitaler Gebäudezwillinge, welche über klassische 3D Visualisierungen (siehe Abbildung 26), hinausgehen. Als Grundlage wird dazu das BRICK-Schema eingesetzt, das auf offenen Semantic-Web-Technologien basiert und eine einheitliche Beschreibung der technischen Gebäudekomponenten ermöglicht. Allerdings erfordert dies ein tieferes Verständnis des zugrunde liegenden Graphen der die Informationen des Gebäudes enthält und der zugehörigen Abfragesprache. Die Nutzung von Sprachmodellen in Kombination mit Retrieval Augmented Generation (RAG) vereinfacht die Arbeit mit dem Gebäudemodell erheblich: In einem ersten Schritt lassen sich hierüber werden die Gebäudekomponenten mittels natürlicher Sprache erfassen und beschreiben. Mittels GenAI können auf Informationsquellen wie Baupläne, Anlagenschaubilder und Datenbanken zugegriffen werden, um eine konsistente Wissensstruktur zu entwickeln und aus diesen zunächst unvollständigen Daten ein präzises digitales Gebäudemodell zu erstellen. Komplexe Abfragen an den Graph können ebenfalls mittels des Sprachmodells in natürlicher Sprache gestellt



Abbildung 26: 3D-Visualisierung eines Gebäudes als Digitaler Zwilling
Quelle: istock.com - sl-f

werden. Der so erstellte Digitale Gebäudezwilling ist äußerst wertvoll, da er neben der Beschreibung des Gebäudes auch für die Optimierung des Betriebs sowie für ein kontinuierliches Monitoring des Gebäudes genutzt werden kann.

Des Weiteren lässt sich solch ein Gebäudemodell in Zukunft auf weitere Anwendungsfälle übertragen. So können beispielsweise Simulationsmodelle zur Wirtschaftlichkeitsbewertung von Umrüstungen erzeugt oder auch Energieausweise erstellt werden. Insgesamt bietet sich die Möglichkeit, durch Sprachmodelle digitale Gebäudemodelle wesentlich einfacher zu erstellen und zu nutzen. Dies bedeutet eine maßgebliche Kompetenzsteigerung in der Gebäudebewirtschaftung, die wiederum genutzt werden kann, um leichter Nachhaltigkeitsziele zu erreichen.

Kontakt

Dr. Christian Kühnert
Fraunhofer IOSB
christian.kuehnert@iosb.fraunhofer.de

4.3.2. Digitaler Zwilling von Fahrzeuginsassen ❷

Moderne Autos sind softwaredefiniert und bieten eine außerordentliche Funktionsvielfalt. Personalisierte Funktionen binden Käufer an eine Marke. Um zudem die neuesten gesetzlichen Vorgaben zur Fahrzeugsicherheit erfüllen zu können, ist ein hohes Maß an Wissen über die Insassen erforderlich. In der Fahrzeugindustrie und bei Fraunhofer wird deshalb an Digitalen Zwillingen der Fahrzeuginsassen gearbeitet.

Diese unterstützen Komfort- und Assistenzsysteme mit Sensorik im Fahrzeuginneren: Sie liefern umfassende Informationen über alle Insassen, ihren aktuellen Zustand sowie ihre Intentionen.

Das Advanced Occupant Monitoring System nutzt optische Sensoren und moderne Maschinelle Lernverfahren, um die 3D-Positionen der Körperteile jedes Fahrzeuginsassen in Echtzeit zu erfassen (siehe Abbildung 27).

Die Erfassung von Körperhaltung und Gestik mit spezialisierten Modellen ist energieeffizient und performant. Für die Interpretation der komplexen Gesamtsituation im Fahrzeuginnenraum werden Vision Language Modelle eingesetzt und mit weiteren multimodalen Daten in multimodalen Modellen ausgewertet. In den Digitalen Zwilling fließen auch Informationen zum Fahrzeugzustand (beispielsweise der Zustand der Automations- und Assistenzsysteme) und über den Kontext (Lichtverhältnisse, Wetter, etc.) ein. Das System ist modular aufgebaut und verwertet mittels Transformer-Modellen nur relevante Informationen für die Anwendung.

Zukünftig werden Digitale Zwillinge der Fahrzeuginsassen wichtig, wenn Fahrerin oder Fahrer nicht permanent selbst lenken müssen und sich dadurch mehr Freiheiten im Fahrzeuginnenraum ergeben. Eine Herausforderung ist in diesem Zusammenhang die Bedienung von Komfort- und Assistenzfunktionen, z. B. über Gesten, statt über weit entfernte Touchscreens. Im Projekt SALSA »Smarte, adaptive und lernbare Systeme für alle« erforscht das Fraunhofer IOSB mit Partnern aus der Industrie, wie Digitale Zwillinge mit Chat-Bots und den Sprachassistenten der Autos kombiniert werden können. Im Forschungsprojekt KARLI »Künstliche Intelligenz

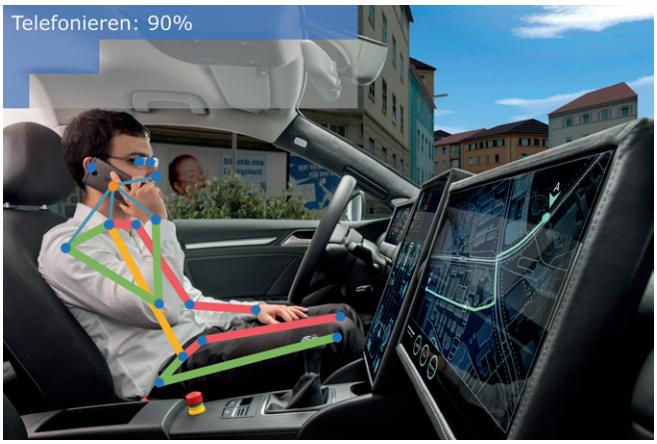


Abbildung 27: Computer Vision Analyse der Fahrzeuginsassen
Quelle: Fraunhofer IOSB

für adaptive, responsive und levelkonforme Interaktionen im Fahrzeug der Zukunft« wurden Insassenaktivitäten für einen Assistenten genutzt, der Kinetose vorbeugt. Hierfür erfasste ein multimodales Model Fahrerzustände in unterschiedlichen Automationsleveln. Zukünftig werden durch Hinzunahme von Sprachanalysen und physiologischen Daten auch kognitive Zustände wie mentale Ablenkung und Schlafphasen im Digitalen Zwilling abgebildet. Im White Paper »Pioneering In-Cabin Monitoring – Unmasking the Power of 2D and 3D Cameras« werden weitere Use Cases beschrieben [41].

Das Advanced Occupant Monitoring System spielt insbesondere in der Vorausentwicklung seine Stärken als modulares und flexibles Forschungssystem aus und liefert eine umfassende Erkennung von Körperhaltung, Gestik, Aktivitäten und der Interaktionen aller Insassen.

Kontakt

Dr. Frederik Diederichs
Fraunhofer IOSB
frederik.diederichs@iosb.fraunhofer.de

4.3.3. Der digitale Produktpass ❸

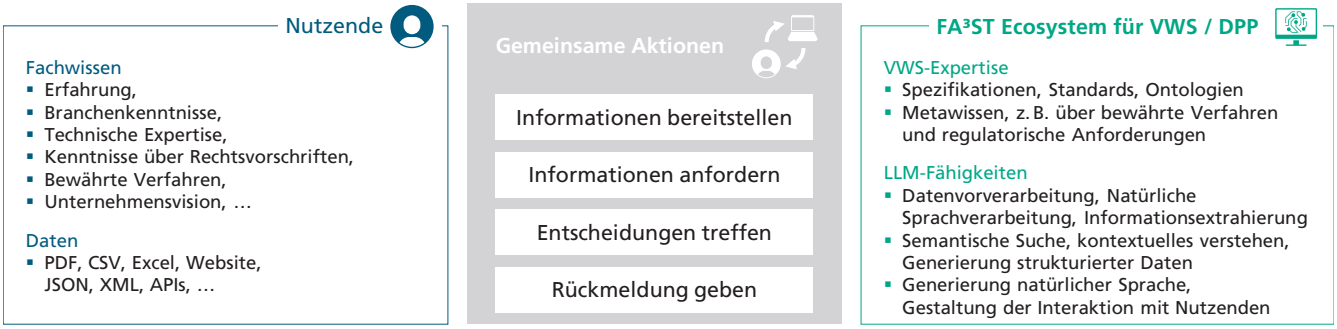


Abbildung 28: Das FA³ST Ecosystem vereint das Domänenwissen und die Daten der Nutzenden mit der VWS-Expertise und LLM-Fähigkeiten.
Quelle: Fraunhofer IOSB

Die Anforderungen an Nachhaltigkeit, Transparenz und Regulierungen in globalen Lieferketten steigen. Durch die EU-Verordnung zur verpflichtenden Einführung Digitaler Produktpässe (DPPs) bis 2027 wird von Unternehmen gefordert, dass sie klare und nachvollziehbare Informationen für den gesamten Produktlebenszyklus bereitstellen.

Ein vielversprechender Ansatz zur Umsetzung von DPPs der Initiative DPP4.0 ist die Nutzung standardisierter Digitaler Zwillinge, die alle relevanten Produktinformationen über standardisierte Schnittstellen bereitstellen.

Ein etablierter Standard hierfür ist die Verwaltungsschale (VWS) (englisch: Asset Administration Shell (AAS)), entwickelt im Rahmen von Industrie 4.0. Die containerartige Struktur mit harmonisierten Teilmodellen lässt sich beispielsweise für technische Daten oder den CO_2-Fußabdruck einsetzen.

VWS ermöglicht die Interoperabilität zwischen verschiedenen Unternehmenssystemen und eine durchgängige Nachverfolgbarkeit von Produktinformationen entlang der gesamten Lieferkette.

Die Umsetzung der VWS ist jedoch komplex und arbeitsintensiv, insbesondere kleine und mittelständische Unternehmen (KMUs) stehen mangels benötigter Ressourcen und Know-hows vor signifikanten Herausforderungen – auch bei der Erfüllung der bevorstehenden regulatorischen Anforderungen.

Um diese Hürden zu überwinden, entwickelt das Fraunhofer IOSB eine intelligente Toolbox: Das FA³ST Ecosystem setzt GenAI in Form von großen Sprachmodellen (LLMs) ein, die speziell für verschiedene Aufgaben nachtrainiert werden und über Retrieval Augmented Generation (RAG) auf VWS-spezifische Informationen zugreifen (siehe Abbildung 28). Ziel ist, Nutzende ohne tiefes VWS-Fachwissen in einem intuitiven Prozess schnell und effizient zu einer standardkonformen VWS zu führen:

Der FA³ST CreAItor erfasst und strukturiert automatisch relevante Daten aus verschiedensten Quellen wie PDFs, Excel-Dateien oder Textdokumenten und konvertiert diese in eine VWS. Der FA³ST ValidAItor prüft die VWS auf Vollständigkeit, Korrektheit und Standardkonformität. Er bietet Hilfestellung in natürlicher Sprache bei der Korrektur von Standardverletzungen – z. B. in Form von Vorschlägen für fehlende Informationen.

Zukünftig können die Nutzenden über den FA³ST Assistant mithilfe von Sprachmodellen Informationen zum DPP in natürlicher Sprache abfragen oder aktualisieren.

Darüber hinaus wird der FA³ST Mapper automatisch die notwendige Konfiguration generieren, um Maschinen bidirektional mit der VWS zu verbinden. Die so gewonnenen kontextuellen Daten können dann für verschiedene Anwendungsfälle verwendet werden, etwa zur präzisen Berechnung des CO_2-Fußabdrucks (PCF) basierend auf realen Verbrauchswerten.

Das Fraunhofer IOSB verfolgt bei diesen Tools einen nutzerzentrierten Ansatz, der die Expertise der Nutzenden einbindet, Standards einhält und die Qualität der DPPs erhöht.

Das ganzheitliche FA³ST Ecosystem für VWS reduziert durch GenAI Zeit, Aufwand, Fachwissen und Personalbedarf für die Umsetzung von DPPs. Intuitive Prozesse vereinfachen die Implementierung von DPPs erheblich und unterstützen insbesondere KMUs dabei, schnell und kostengünstig gesetzliche Anforderungen zu erfüllen.

Kontakt

Dr. Ina Steinmetz
Fraunhofer IOSB
ina.steinmetz@iosb.fraunhofer.de

4.3.4. Semantisches Datenmapping bei Digitalen Zwillingen ❹

Digitale Zwillinge von industriellen Maschinen aufzubauen, erfordert eine präzise Verknüpfung von realen Maschinendaten mit strukturierten Informationsmodellen des Zwillings. Eine zentrale Herausforderung dabei ist die semantische Datenverknüpfung, insbesondere wenn Daten über OPC UA (Open Platform Communications Unified Architecture) bereitgestellt werden. Die OPC-UA-Datenmodelle sind in der Regel baumartig organisiert und bestehen aus zahlreichen Knoten, die Parameter wie Temperatur, Druck, Statusinformationen oder Alarme repräsentieren. Diese Knoten sind jedoch oft uneinheitlich benannt, inkonsistent dokumentiert oder herstellerspezifisch kodiert, was eine automatisierte Interpretation erschwert. Gleichzeitig verlangen standardisierte Digitale Zwillinge – etwa gemäß der Verwaltungsschale (VWS) (englisch: Asset Administration Shell (AAS)) – eine klare, semantisch eindeutige Zuordnung dieser Datenpunkte zu definierten Submodellen und Eigenschaften.

Um die semantische Datenverknüpfung zwischen den OPC UA-Datenknoten eines Assets und den standardisierten Modellen des Digitalen Zwillings zu automatisieren, muss daher eine Integration und Zuordnung von Daten aus unterschiedlichen Quellen stattfinden, um eine konsistente und aktuelle Darstellung des physischen Objekts oder Systems zu gewährleisten.

GenAI, insbesondere Sprachmodelle (LLMs), bieten hier einen vielversprechenden Lösungsansatz. Sie können als semantische Vermittler agieren, indem sie die Bedeutung von OPC-UA-Knoten analysieren und automatisch mit den entsprechenden Strukturen im Digitalen Zwilling abgleichen.

Sprachmodelle (LLMs) können verwendet werden, um automatisch relevante Daten aus den OPC-UA-Knoten zu extrahieren und zu interpretieren. Dazu werden Metadaten wie Knotenname, Einheit, Beschreibung und Kontextinformationen als Eingabe verwendet. Das Modell interpretiert diese Informationen und schlägt eine passende Zuordnung zu einem standardisierten Zwillingsmodell vor.

Sprachmodelle (LLMs) analysieren die Bedeutung der OPC-UA-Knoten und ordnen sie den Submodellen und Eigenschaften im Digitalen Zwilling semantisch zu. Ein Knoten mit der Bezeichnung »AI01.Temp1« und der Beschreibung »Kühlwassertemperatur« kann z. B. automatisch als Temperature.InletCoolingWater



Abbildung 29: Semantisches Datenmapping in der Umsetzung
Quelle: Fraunhofer FFB

in einem Submodell »Environment« erkannt und zugeordnet werden.

Diese Prozesse ermöglichen eine erhebliche Zeitersparnis gegenüber der manuellen Pflege und sind skalierbar über unterschiedlichste Maschinen und Datenquellen hinweg. Durch menschliches Feedback kann der Prozess iterativ verbessert werden, um die Genauigkeit und Zuverlässigkeit der Zuordnungen zu erhöhen. Die generierten Zuordnungen lassen sich strukturiert exportieren und direkt in bestehende Digitale-Zwilling-Plattformen integrieren. Dies sorgt effizient, konsistent und zukunftssicher für die semantische Interoperabilität zwischen realer Anlage und digitalem Abbild.

Durch den Einsatz von Sprachmodellen (LLMs) können Unternehmen den Aufwand für die manuelle Datenverknüpfung reduzieren, die Effizienz steigern und die Genauigkeit der Digitalen Zwillinge verbessern. Dies ermöglicht eine nahtlose Integration und Verwaltung von Daten aus verschiedenen Quellen und trägt zur Optimierung industrieller Prozesse bei.

Kontakt

Dr. Jonathan Krauß
Fraunhofer FFB
jonathan.krauss@ffb.fraunhofer.de

5. Multimodale Daten und Robotik

5.1. Die Datenstruktur

In modernen Roboter- und Anlagensystemen arbeiten zahlreiche Sensoren parallel. Sie sorgen für eine ganzheitliche Wahrnehmung der Umgebung und ermöglichen dem Roboter in Echtzeit passend zu reagieren.

Die wichtigsten Quellen sind meistens Kameras, die sowohl RGB-Farbbilder als auch Entfernungsdaten in variablen FPS (Bilder pro Sekunde) liefern. Ergänzt werden sie durch 3D-Laser-Scanner wie LiDAR oder Radar, die dichte Punktwolken für Entfernungen und Formen erzeugen. Mikrofone erfassen Maschinengeräusche, Sprache oder Warnsignale. Mechanische Sensoren – Kraft- und Drehmomentaufnehmer, IMU-Einheiten (Inertial Measurement Units, messen Beschleunigung und Drehgeschwindigkeit) sowie Temperaturfühler – runden das Sensorspektrum ab.

Erst die kombinierte Auswertung all dieser Quellen liefert ein detailreiches Umgebungsmodell, doch gerade diese Fusion ist technisch anspruchsvoll: Datenraten, Formate und Blickwinkel müssen permanent aufeinander abgestimmt werden. Selbst in einem sorgfältig abgeleiteten Umgebungsmodell bleiben die Daten heterogen und schwer überschaubar, was es schwierig macht, sie für eine präzise Steuerung des Roboters zu nutzen. Erschwerend kommen dynamische Einsatzorte und wechselnde Aufgaben hinzu. In der Produktion führen beispielsweise ständige Produktänderungen oder kleine Stückzahlen dazu, dass Roboterprogramme regelmäßig angepasst werden müssten – ein Prozess, der häufig zu zeit- und geldaufwendig ist und somit eine Automatisierung verhindert.

Die Robotersteuerung basiert aktuell auf vordefinierten Umgebungsmodellen. Diese können zwar automatisch erzeugt werden, müssen in der Regel aber manuell angereichert werden, damit Befehle und Interaktionen ausgeführt werden können. Sobald die Steuerung von Robotern nicht nur auf vordefinierten Umgebungsmodellen arbeitet, sondern in der Lage ist, die Umgebung zu interpretieren und darauf einzugehen, ergeben sich zahlreiche Nutzungsszenarien. In der Mensch-Roboter-Interaktion (HMI) lassen sich sprachgesteuerte Befehle direkt mit Kamerabildern abgleichen und Fehlinterpretationen oder Sicherheitsrisiken erkennen, damit die Anlage bei Bedarf automatisch stoppt. Auf Baustellen lassen sich Bagger dank des robusten Situationsbewusstseins sicher und autonom durch unwegsames Gelände manövrieren.

Die Flexibilität dieser Ansätze erlaubt es auch, dass Roboter sicher in von Menschen genutzten Räumen operieren, wie z. B. beim autonomen Kommissionieren in Lagerhäusern.

5.2. Warum GenAI für diese Datenstruktur?

Die Integration von Robotik und Künstlicher Intelligenz steht vor komplexen Herausforderungen, die durch den Einsatz moderner Foundation-Modelle zunehmend bewältigt werden können. Die Robotersteuerung in realen Umgebungen erfordert die Verarbeitung multimodaler Informationen, die in ihrer Rohform heterogen und uneinheitlich bleiben. Ein zentrales Problem besteht darin, dass Objekte und Situationen semantisch erkannt werden müssen – beispielsweise erfordert ein Mensch auf der Fahrstrecke eines Roboters ein anderes Verhalten als ein statisches Hindernis wie ein Karton, und eine Tasse muss anders gegriffen werden als eine Pflanze. Flexibles Verhalten in komplexen Szenarien setzt somit ein echtes »Verständnis« der Umgebung voraus. Eine zusätzliche Schwierigkeit liegt in der unterschiedlichen Kodierung der Sensorströme (Bild, Ton, Kraft), was die direkte Ableitung von Aktionen erschwert.

Foundation-Modelle bieten mehrere entscheidende Vorteile für diese Herausforderungen. Sie ermöglichen ein semantisches Weltverständnis, bei dem das Modell kontextbewusst zwischen verschiedenen Objekten unterscheiden kann und dieses Wissen direkt in das Roboterverhalten einfließen lässt. Ein gemeinsamer multimodaler Embedding-Raum führt Bild-, Audio- und Sensordaten in einer einheitlichen Repräsentation zusammen, wodurch sie konsistent ausgewertet und in Aktionen umgesetzt werden können.

Besonders wertvoll ist die Zero- bzw. Few-Shot-Fähigkeit der Foundation-Modelle, die den Umrüst- und Trainingsaufwand erheblich reduziert. So können durch Foundation-Modelle Aufgaben ohne oder mit nur wenigen Beispielen ausgeführt werden, ohne explizit für diese trainiert worden zu sein. Diese Fähigkeiten lassen sie schnell auf neue Anforderungen reagieren und ohne langwieriges manuelles Erzeugen und Bereinigen von Trainingsdaten einsetzen. Zudem bietet der Zugriff auf

vortrainiertes Hintergrundwissen robuste Kontextschlüsse auch bei neuen Aufgaben.

Die Einsatzmöglichkeiten von Foundation-Modellen in Robotik und Automatisierung sind vielfältig. Im Bereich der Programmierung ermöglichen sie die Erzeugung von Programmen und Abläufen unter anderem für Programmiersprachen zur Prozessautomatisierung aus simplen Beschreibungen (z. B. für SPS oder Behavior Trees). Für die Kontrolle von Robotern werden Vision-Language-Action-Modelle mit einer Kombination aus visueller, sprachlicher und aktionsbedingter Eingabe und Ausgabe trainiert, die Aufgaben wie Greifpunktvorhersage, Objektposen-Schätzung und sogar die direkte Kontrolle über den Roboter übernehmen können. Vision-Language Models (VLMs) finden Anwendung in der Texterkennung (OCR) und Dokumentenverarbeitung. Hier werden Bild und Textdaten analysiert und Text erzeugt. Dies trägt maßgeblich zu dem Auflösen von Barrieren wie Sprachdifferenzen oder Sehen bei. VLMs unterstützen außerdem die Re-Identification (Re-ID) für Produktabgleich und automatische Beschreibungen. Die Kombination aus Text und Bildeingabe schafft die Basis für Modelle, wie Segment Anything, DINOv2 und Depth Anything. Diese ermöglichen das Finden und Segmentieren von Objekten in Bildern und das Schätzen von Tiefeninformationen von RGB-Sensoren ohne die Notwendigkeit von einer großen Menge an Trainingsdaten. Sogenannte Open Vocabulary Detektoren (Grounding Dino, DETR) ermöglichen ohne weiteres Training semantisches Kategorisieren von ganzen Szenen, in denen jedes einzelne Objekt in der Szene mit einem Begrenzungsrahmen (Bounding Box) versehen und gelabelt wird. In der Aufgabenplanung eröffnen hierarchische Ende-zu-Ende-Ansätze mit Agenten neue Möglichkeiten, die dem menschlichen Planen ähneln. Die Agenten können zusammenarbeiten, delegieren, diskutieren und so viele komplexe Aufgaben bewältigen.

Trotz dieser Fortschritte bestehen noch Grenzen und offene Probleme. Die Präzision bei feinkörniger Posenerkennung (Pose Estimation) ist nach wie vor eingeschränkt. Halluzinationen und Über-Konfidenz (falsche Sicherheit) bei fehlenden Modalitäten stellen Herausforderungen dar. Diese beiden Faktoren können die Zuverlässigkeit und Vertrauenswürdigkeit von Ergebnissen und Entscheidungen beeinträchtigen. Der hohe Rechen- und Energiebedarf für die Ausführung solcher Modelle direkt auf Edge-Geräten (lokale, kleine Verarbeitungseinheiten) anstatt auf großen Recheneinheiten bleibt ein limitierender Faktor. Nicht zuletzt erfordern Datenschutz und die Zertifizierung sicherheitskritischer FM-Entscheidungen besondere Aufmerksamkeit und Lösungsansätze für den industriellen Einsatz.

Tabelle 6: Multimodale Daten und Robotik – Uses Cases in der Übersicht

Branche/ Bereich	Use Case	Aufgabentyp	Datentyp & Datenstruktur	Modell oder Methode
Bau	❶ Steuerung von autonomen Baumaschinen	Autonome Baumaschinen mit natürlicher Sprache steuern	Sensordaten (RGB-D, LiDAR, IMU, GPS), Sprachanweisungen, Aktionen	LLMs, VLMs
Industrie & Produktion	❷ Robotik & Anlagenprozesse mit Sprach- & KI-Agenten steuern	Roboter und Anlagen mit natürlicher Sprache steuern	Sensor- und Kameradaten (RGB-D, LiDAR), Sprachanweisungen	KI-Agentensystem mit Zugriff auf weitere LLMs, VLMs, Vision Foundation-Modelle und den entsprechenden Roboter/die entsprechende Anlage
Produktion	❸ Kollaborative Interaktion mit Robotern in der Produktion	Mitarbeitende durch KI gestützte Steuerung von Robotern sicher und effizient unterstützen	Sensordaten (RGB-D), Sprachanweisungen	LLMs, VLMs, VLAs, Pose- und Gaze Merkmalextraktor
Logistik	❹ Revolutionierung der Kommissionierung: KI für die Logistik	Kommissionierung automatisieren, Objekte greifen, Lagerhäuser navigieren	Sensordaten (RGB-D, Lidar, IMU), Sprachanweisungen, Dokumente, Umgebungskarten	LLMs, VLMs, Vision Foundation Modelle, Reinforcement Learning
Re-Commerce, Retouren	❺ Intelligente Produktbeschreibung bei Retouren und Re-Commerce	Produkte identifizieren, Stammdaten für den Re-Commerce prüfen und automatisch generieren	Bilddaten (RGB-D), Datenbank (Stammdaten), Menschliche Eingaben	Re-Identification, LLMs, VLMs

5.3. Use Cases

5.3.1. Steuerung von autonomen Baumaschinen 1

Die Steuerung autonomer Baumaschinen (siehe Abbildung 30) ist eine der anspruchsvollsten Aufgaben in der Robotik, da Maschinen Objekte, Gelände und Gefahren erkannt und in Echtzeit sichere, effiziente Entscheidungen getroffen werden müssen. Bisherige modulare Systeme analysieren Sensordaten separat und steuern Roboterverhalten auf Basis vorgegebener Handlungspläne, was bei Generalisierbarkeit und Robustheit an Grenzen stößt. Multimodale Foundation-Modelle bieten hier ein disruptives Potenzial.

Ein besonders relevanter Anwendungsfall liegt in der kontextsensitiven Aufgabenplanung und der Handlungsentscheidung von autonomen Baumaschinen, wie z. B. Baggern, die in kontaminierten Gebieten selbstständig Erdschichten abtragen oder auf Baustellen Oberflächen modellieren. In der Praxis bedeutet dies, dass die Maschinen nicht nur unterschiedliche Sensordatenströme (Kamera, LiDAR, IMU, GPS etc.) integrieren müssen, sondern auch natürliche Sprache (z. B. Einsatzanweisungen) oder Pläne verarbeiten und in sichere Handlungen umsetzen müssen. Ein typisches Beispiel: Das Bedienpersonal beschreibt verbal ein Ziel wie »entferne die oberste Erdschicht bis zur Tiefe von 30 Zentimetern innerhalb dieses Bereichs«, während die Maschine vor Ort Hindernisse wie Steine oder kontaminierte Bereiche erkennt, um die beste Vorgehensweise abzuleiten.

Multimodale Foundation-Modelle, die durch große Mengen domänenspezifischer Daten (Bilddaten, Punktwolken, Sprache, Aktionsprotokolle) trainiert wurden, besitzen die Fähigkeit, über Modalitäten hinweg Bedeutungen in einem einheitlichen semantischen Raum zu verknüpfen: visuelle Sensorinformationen mit sprachlichen Aufgabenbeschreibungen, taktile Rückmeldungen mit planerischen Kontexten. So können durch Prompting situationsspezifische Anweisungen in robotische Handlungsketten überführt werden – selbst bei bisher ungesehenen Aufgabenstellungen. Gerade in sich verändernden Umgebungen wie Baustellen kann diese Fähigkeit zur situationsangepassten Interpretation und Reaktion einen entscheidenden Effizienzgewinn bringen.

Der erwartete Nutzen multimodaler Foundation-Modelle ist erheblich: Sie ermöglichen nicht nur eine bessere



Abbildung 30: Autonomer Bagger des Fraunhofer IOSB beim Beladen eines Brechers
Quelle: Fraunhofer IOSB

Generalisierbarkeit und eine höhere Robustheit gegenüber Sensorstörungen, sondern auch eine bessere Skalierbarkeit auf unterschiedliche Maschinenplattformen sowie eine kontinuierliche Verbesserung durch neue Datenquellen. Um dieses Potenzial voll auszuschöpfen, wird derzeit an einer skalierbaren Systemarchitektur gearbeitet, die domänenübergreifendes Training mit realen Einsatzdaten, Simulationen und kollaborativ gesammelten Demonstrationen kombiniert. Besonders vielversprechend ist dabei die Integration in bestehende Flottenmanagementsysteme und die Nutzung von Simulationsumgebungen für Pretraining und Szenarientests. Perspektivisch können multimodale Foundation-Modelle so nicht nur einzelne Maschinen steuern, sondern als zentrale, lernfähige Wissensschicht zwischen Mensch, Sensorik und Aktorik agieren – adaptiv, generalisierbar und übergreifend über Aufgaben und Prozesse hinweg.

Kontakt

Dr.-Ing. Janko Petereit
Fraunhofer IOSB
janko.petereit@iosb.fraunhofer.de

5.3.2. Robotik und Anlagenprozesse mit Sprach- und KI-Agenten steuern ❷

Industrieroboter und -anlagen werden heute mit einer Vielzahl heterogener Programmiersprachen, von SPS über Behavior Trees bis zu Python-Skripten, eingerichtet. Jede Umrüstung zum Beispiel auf ein neues Produkt bedeutet daher zeitraubende Anpassungen und hohe Engineering-Kosten. Gerade mittelständische Betriebe mit kleinen Losgrößen, Saison- oder Verpackungsvarianten leiden unter langen Inbetriebnahmezeiten, die ihre Wirtschaftlichkeit ausbremsen oder eine Automatisierung unmöglich machen.

Komplexe Sequenzen, Fehler und Ausnahmeroutinen müssen bislang händisch modelliert und programmiert werden. Das verlangt seltene Expertise und macht Unternehmen von externem Know how abhängig. Fehlen Expertinnen und Experten, bleibt die Anlage stehen – mit entsprechenden Kosten für Stillstand und Terminverzug.

Foundation-Modelle können natürliche Sprache in semantische Roboterbefehle übersetzen: Ein Satz wie »Palettiere zehn Kartons auf Palette B und melde nach jedem Layer, ob das Muster stimmt« wird automatisch in einen generierten Ablauf und eine Bewegungslogik übersetzt. Um diese Fähigkeit abzubilden, wird das multimodale Weltverständnis solcher Modelle mit Kamerabildern, Kraft-Sensoren oder IoT-Daten ergänzt, damit der Roboter kontextdynamisch den Befehl ausführen, reagieren und interpretieren kann. Das entwickelte KI-Agenten-Framework ermöglicht es, die Aufgabe in Teilaufgaben zu zerlegen, Programmteile bzw. Bewegungspläne zu generieren, Monitoring-Agenten (z. B. VLMs) die Durchführung überwachen zu lassen und bei Abweichung zu optimieren oder zu stoppen. So entsteht ein selbstkorrigierender, dialogfähiger Ablauf, der vom Menschen initiierte Aufgaben intelligent für Roboter übersetzt, zerlegt, kontinuierlich überwacht und dynamisch optimiert.

Der Vorteil: Die Beschreibung der Routineabfolgen inklusive Fehler und Ausnahmebehandlung entstehen per Prompt, anstatt manuell per Hand programmiert zu werden. Das Bedienpersonal kann während des Betriebs per Sprache mit dem Roboter interagieren und Neues oder Veränderungen mitgeben. Zum Beispiel: »Setze nach dem Greifen eine Verzögerung von 200 Millisekunden, falls die Kamera Vibration erkennt.« Das Feedback des Bedienpersonals fließt als Lernsignal zurück, sodass das System kontinuierlich seine Routinen verfeinert.



Abbildung 31: Fraunhofer IAIS Demonstrator – ein per Sprache programmierbarer Roboterarm
Quelle: Fraunhofer IAIS

In dem in Abbildung 31 dargestellten Fraunhofer-Demonstrator interagieren zum Beispiel ein kollaborativer Roboterarm, eine RGBD-Kamera und eine webbasierte Sprach-GUI. Für typische Pick&Place-Aufgaben sank die Inbetriebnahmezeit deutlich, da der Programmcode dialogbasiert generiert und iterativ optimiert werden konnte.

Unternehmen profitieren von GenAI durch den Übergang zu Robotik, die wenig oder keine Programmierung erfordert, was die Einstiegshürde senkt und schnelle Variantenwechsel ohne tiefes Programmierwissen ermöglicht. Dies führt zu kürzeren Rüst- und Stillstandszeiten und macht eine wertschöpfende Automatisierung auch für kleine Serien wirtschaftlich realisierbar. Dieses KI-Agenten-Framework bietet somit einen skalierbaren Weg zu wandlungsfähigen Produktionszellen, in denen Sprachbefehle genügen, um neue Abläufe sicher, effizient und nachvollziehbar zu implementieren, und kontinuierlich zu optimieren.

Kontakt

Alexander Zorn
Fraunhofer IAIS
alexander.zorn@iais.fraunhofer.de

5.3.3. Kollaborative Robotik mit Pose- und Gaze-Tracking 3

In modernen Produktionsstätten gewinnen kollaborative Roboter (»Cobots«) zunehmend an Bedeutung. Im Gegensatz zu traditionellen Industrierobotern, die üblicherweise in geschützten Sicherheitsbereichen eingesetzt werden, interagieren Cobots aktiv mit menschlichen Mitarbeitenden im selben Arbeitsbereich und unterstützen diese bei anspruchsvollen Aufgaben.

Um eine sichere, effiziente und reibungslose Zusammenarbeit zu gewährleisten, muss der Roboter seine Umgebung visuell erfassen, gesprochene Anweisungen verstehen sowie die Absichten und Handlungen seiner menschlichen Kollegen antizipieren und interpretieren können. Ein wesentlicher Faktor hierfür ist der Einsatz zusätzlicher Sensorik, insbesondere Gaze-Tracking und Pose-Tracking (siehe Abbildung 32). Gaze-Tracking bezeichnet Technologien zur präzisen Verfolgung der Blickrichtung einer Person. Kameras oder spezielle Sensoren erfassen die Augenbewegungen und ermitteln, welches Objekt, Werkzeug oder welcher Arbeitsbereich gerade im Fokus steht. Pose-Tracking hingegen analysiert mithilfe von Kameras oder Wearables die Körperhaltung und -bewegung des Menschen. Das System erkennt beispielsweise, ob eine Person steht, sich beugt, einen Arm hebt, eine Drehbewegung ausführt oder sich einem Objekt nähert.

Diese Sensordaten liefern nicht nur eine Momentaufnahme der menschlichen Aktivität, sondern bilden auch die Grundlage für die Absichtserkennung. Wenn ein bestimmtes Bauteil längere Zeit betrachtet wird, kann der Roboter davon ausgehen, dass dieses als Nächstes bearbeitet werden soll. Hebt der Mensch die Hand in Richtung eines Werkzeugs oder nähert er sich dem Arbeitsbereich des Roboters, kann der Roboter rechtzeitig reagieren, beispielsweise durch Verlangsamung, Anhalten oder gezielte Unterstützung. Die Daten aus Gaze- und Pose-Tracking werden zusammen mit Bild- und Sprachdaten in Vision-Language-Action-Modelle integriert, um ein umfassendes, multimodales Situationsverständnis zu erzeugen. Diese Modelle können nicht nur auf sprachliche Anweisungen (»Reiche mir den Schraubenzieher«) und visuelle Hinweise reagieren, sondern auch den nächsten Arbeitsschritt antizipieren, indem sie die Blickrichtung und Körperhaltung des Menschen interpretieren. In der Praxis bedeutet dies zum Beispiel:



Abbildung 32: Pose und Blickrichtung des Menschen wird erkannt während der Kollaboration mit einem Roboter. Quelle: Fraunhofer IPA

- Bei der gemeinsamen Montage erkennt der Roboter, wann er ein Bauteil bereithalten oder den Arbeitsplatz sichern muss.
- Er kann vorhersehen, welche Teile als Nächstes benötigt werden, und diese griffbereit positionieren.
- Bei potenziellen Kollisionen kann das System sofort stoppen oder ausweichen, noch bevor eine gefährliche Situation entsteht.

Durch die Integration von Gaze- und Pose-Tracking wird der Roboter zu einem proaktiven und aufmerksamen Teamplayer. Er versteht nicht nur die gesprochenen Anweisungen, sondern auch die Bewegungen eines Menschen und kann sein Verhalten dynamisch anpassen. Dies erhöht die Sicherheit im gemeinsamen Arbeitsbereich, reduziert Fehler und gestaltet die Interaktion insgesamt natürlicher und effizienter.

Kontakt

Dr. Florian Strohm
Fraunhofer IPA
florian.strohm@ipa.fraunhofer.de

5.3.4. Revolutionierung der Kommissionierung: GenAI für die Logistik 4

In modernen Logistikzentren steigt der Druck, Prozesse immer weiter zu automatisieren und zugleich flexibel auf wechselnde Rahmenbedingungen zu reagieren. Insbesondere in »Brownfield«-Szenarien, wo autonome Systeme sich ihren Arbeitsraum mit menschlichen Arbeitskräften und konventionellen Fördermitteln teilen, sind präzise Wahrnehmung und kontinuierliche Umfeldanalyse entscheidend.

Ein vielversprechender Ansatz dafür ist der Einsatz multimodaler GenAI zur automatisierten Kommissionierung. Im ersten Schritt der automatisierten Kommissionierung erkundet der mobile Roboter (siehe Abbildung 33) seine Umgebung und ergänzt das klassische SLAM (Simultaneous Localization And Mapping)-Verfahren, in dem der Roboter eine Karte zur Selbstlokalisierung generiert, um semantische Metadaten. Dafür werden ausschließlich jene Sensordaten in das Vision-Language-Modell (VLM) eingespeist, die das Modell zuverlässig verarbeiten kann: hochauflösende RGB-Kameraaufnahmen sowie Tiefen- und Lidar-Punktewolken. Auf Basis dieser drei Modalitäten erzeugt das VLM Beschreibungen von Regalen, Gängen und Lagerbehältern (z. B. »Regal A3, Fach 2, Karton mit Schrauben«), die für den späteren Gebrauch in den Datenstrukturen (Graphen oder Datenbanken) abgelegt werden und dann vom Robotersystem genutzt werden können, um Navigations- und Kommissionierbefehle abzuleiten. Zusätzlich ist es in der Lage, durch Funktionsaufruf die richtigen Roboterfähigkeiten wie »Greife Objekt x« oder »Navigiere zu Regal y« in der richtigen Reihenfolge anzuwenden. Durch fortlaufende »Semantic Mapping«-Zyklen (Semantische Kartierung) entsteht ein dynamisch aktualisiertes Modell der Umgebung.

Am Zielregal kann nun eine weitere Art von Foundation-Modellen benutzt werden: Objektposen-Schätzer. Modelle zur Positions- und Orientierungserkennung können anhand weniger Referenzbilder oder eines 3D-Modells des gewünschten Objekts in Echtzeit aus einem Kamera-Stream Objektposen nachverfolgen. Damit wird der Greifvorgang für den Roboter stark vereinfacht. Nun analysiert das Gesamtsystem ausgehend von der Kommissionieranfrage, die durch Sprachbefehle



Abbildung 33: Mobiler Manipulatoren-Prototyp des Fraunhofer IML: mögliche Plattform für den Kommissionierungs-Agenten
Quelle: Sebastian Hoose

(z. B. »Zehn Schrauben aus Fach B4«) oder durch Analyse von Dokumenten erfolgt, die aktuellen Kamerabilder und Tiefendaten, führt eine Anfrage an die Bildsegmentierung und Objektposenschätzung durch und leitet diese Informationen an die nachfolgende Greifplanung weiter. Diese kann entweder auf heuristischen Strategien und Planung basieren oder durch lernende Verfahren, etwa durch Reinforcement Learning oder Imitation Learning, adaptiv gestaltet werden. So wird die Greifaktion flexibel an das jeweilige Objekt und dessen Kontext angepasst – auch dann, wenn es sich um zuvor unbekannte Artikel handelt.

Auf diese Weise wird die gesamte Kommissionierkette – von der initialen Umfelderkundung bis zur Greifaktionsplanung – durch multimodale GenAI robust und effizient gestaltet und kann in unbekannten Umgebungen eingesetzt werden.

Kontakt

Nicolas Bach
Fraunhofer IML
nicolas.bach@iml.fraunhofer.de

5.3.5. Intelligente Produktbeschreibung bei Retouren und Re-Commerce 5

In der Retouren- und Re-Commerce-Logistik stoßen traditionelle Methoden zur Produktidentifikation, wie Barcodes oder RFID, oft an Grenzen. Entgegen dem Idealbild, wie in Abbildung 34 dargestellt, fehlen diese Identifikatoren. Besonders bei beschädigten, gebrauchten oder unvollständig gekennzeichneten Artikeln entstehen Fehler und Verzögerungen bei der Zuordnung und Wiedervermarktung. Dadurch ergibt sich ein hohes Maß an qualitativen Artikeln, die entsorgt werden, weil die Kosten zur Identifikation, Aufbereitung und zum Erstellen von Produktbeschreibungen höher als der Wiederverkaufswert sind.

Eine innovative Lösung bietet ein System mit dem Einsatz von multimodalen Modellen. Diese nutzen eine Kombination aus Sprach-, Bild- und Textanalyse, um Produkte auch dann präzise zu identifizieren, wenn Produktbeschreibungen nicht mehr vorhanden sowie klassische Merkmale beschädigt sind oder fehlen. Der Prozess funktioniert in vier Schritten:

1. **Bilderkennung:** Computer-Vision-Modelle analysieren Produktfotos und erkennen zentrale Merkmale wie Form, Farbe oder Markenlogos. Insbesondere bei der Farb- und Formerkennung können VLMs präzise Informationen aus Bilddaten extrahieren. Bei branchenspezifischen oder neuen Marken kann durch fehlende Repräsentation in den Trainingsdaten eine zuverlässige Marken-Zuordnung nicht sichergestellt werden. Durch Re-Identifikations-Modelle können Logos aus unterschiedlichen Perspektiven zuverlässig identifiziert werden, indem sie eine Vektordarstellung erzeugen und mit bekannten abgleichen.

2. **Automatische Texterkennung und sprachliche Kontextanalyse:** Das System liest und interpretiert sichtbare Textelemente, auch wenn diese beschädigt oder unvollständig sind. Zusätzlich berücksichtigt es bei Re-Commerce sprachliche Informationen aus von Endnutzerinnen und -nutzern gegebenen Beschreibungen, um Hinweise wie Hersteller, Modellnummer oder Produktmerkmale abzuleiten.

3. **Multimodaler Abgleich:** Diese visuellen und textbasierten Informationen fließen in eine digitale »Signatur« des Artikels ein. Diese Signatur wird anschließend mit einem Produktkatalog abgeglichen, der neben Bildern auch durch VLMs angereicherte bzw. erzeugte detaillierte sprachliche Beschreibungen enthält. Hierdurch können fehlende Stammdaten für bekannte Artikel wieder verfügbar gemacht werden.



Abbildung 34: Scan einer Retoure zum Versand
Quelle: Grustock - stock.adobe.com

4. **Automatische Produktbeschreibung:** Basierend auf diesen visuellen und sprachlichen Daten generiert ein VLM aussagekräftige und zuverlässige Produktbeschreibungen für den Wiederverkauf. Die Multimodalität erhöht hierbei Genauigkeit und Vertrauenswürdigkeit, auch wenn Informationen unvollständig oder komplex sind.

Gebrauchte Artikel werden automatisch und schnell klassifiziert. Dies beschleunigt die Wiedereinführung neuwertiger Artikel in den Markt und reduziert manuelle Eingriffe und Fehler deutlich und spart Kosten in der Logistik.

Auch für die Kundinnen und Kunden ergeben sich Vorteile: Rückerstattungen erfolgen schneller, und Produktbeschreibungen bei Secondhand-Artikeln werden zuverlässiger. So wird die Bereitschaft, nicht mehr gebrauchte Artikel wieder zu verkaufen, erhöht. Dieser Ansatz trägt zu mehr Nachhaltigkeit bei, da deutlich weniger Ware entsorgt und mehr in den Markt zurückgeführt wird.

Kontakt

Christian Pionzewski
Fraunhofer IML
christian.pionzewski@iml.fraunhofer.de

6. Fazit und Ausblick

6.1. Management Summary

In diesem Whitepaper wurden Chancen aufgezeigt und Möglichkeiten illustriert, wie GenAI-Modelle, bekannt für ihre Fähigkeiten in Bezug auf Text, Audio, Bild und Video, auch auf komplexe Daten in den Kernprozessen von Unternehmen angewendet werden können. Dazu wurden für verschiedene Use Cases aus den unterschiedlichsten Branchen Modelle nachtrainiert, spezialisierte Modelle genutzt, Prompting-Methoden verfeinert, RAGs integriert oder Modelle miteinander kombiniert. Die Use Cases zeigen eine Vielzahl signifikanter Vorteile gegenüber älteren Technologien:

- Keine Technologie vor GenAI konnte verschiedenartige Daten so umfassend verstehen und verarbeiten. Erstmals ist es möglich, Informationen aus komplexen, ursprünglich für Menschen bestimmten Quellen sowie aus einer Vielzahl paralleler Daten zu extrahieren und unmittelbar weiterzuverarbeiten.

- GenAI-Modelle können umfangreiche und komplexe Daten in einer einzigen Anfrage verarbeiten und dabei selbst in komplexen Kontexten Zusammenhänge und Muster zu erkennen.

- Frühere datenbasierte Modelle erforderten große Mengen gut gepflegter und annotierter Daten. Dies stellte oft eine erhebliche Herausforderung dar, die häufig nicht angegangen wurde. Im Gegensatz dazu sind GenAI-Grundmodelle vortrainiert und können bereits durch geschicktes Prompting eingesetzt werden. Dadurch lassen sich erste Experimente nahezu ohne eigene Daten durchführen. Zudem gibt es etablierte Verfahren, um Grundmodelle zu spezialisieren, wofür vergleichsweise wenige Unternehmensdaten benötigt werden. Der Einstieg in GenAI-Lösungen ist daher unkompliziert und leicht.

- GenAI-Modelle können nicht nur Daten verstehen und analysieren, sie können auch neuartige Inhalte generieren. In diesem Whitepaper wurde diese Fähigkeit etwa genutzt, um zusätzliche Trainingsdaten zu generieren oder eigene Daten zu ergänzen, zu korrigieren oder zu transformieren.

- GenAI-Modelle sind in der Lage, Aufgaben zu bewältigen, die ihnen in natürlicher Sprache übermittelt werden, ohne dass sie diese zuvor erlernt haben müssen. Dies ermöglicht auch Personen ohne Fachwissen, die Modelle direkt zu bedienen und in neuen Situationen innerhalb ihres Arbeitskontexts einzusetzen. Programmierkenntnisse sind nicht erforderlich. Diese Fähigkeit, komplexe Analysen ohne tiefgehendes Fachwissen durchzuführen, demokratisiert den Zugang zu datenbasierten Entscheidungen und fördert Innovationen in verschiedenen Branchen.

- Weitere Beispiele verdeutlichen anschaulich, wie GenAI-Modelle Wissen generieren können, das anschließend in Form von Wissensgraphen als gesicherte Hintergrundinformation bereitgestellt werden kann.

- Im Bereich der Materialwissenschaften, Chemie und Pharmazie finden wir Beispiele, wo GenAI-Modelle ganz neue, nie dagewesene Molekülstrukturen mit gewünschten Eigenschaften generieren können.

Für fünf einschlägige Datentypen aus Unternehmenskontexten hat dieses Whitepaper gezeigt, wie GenAI bestehende Verfahren ersetzen kann, welche Mehrwerte dies bietet und wie GenAI ganz neue Möglichkeiten schafft, um die Produktivität zu steigern. Mit GenAI können Unternehmen bekannte Aufgaben effizienter lösen oder ganz neue Fragestellungen angehen und dabei qualitativ hochwertige Erkenntnisse gewinnen:

1. **Tabellarische Daten:** Spezialisierte Modelle für tabellarische Daten verstehen die semantischen Bedeutungen von Werten im Kontext ihrer Verwendung. Dies erhöht die Effizienz der Datenverarbeitung und reduziert den Aufwand für die Kontextgenerierung deutlich. So lassen sich schnellere und kosteneffizientere Analysen erstellen, die Unternehmen helfen, datengetriebene Entscheidungen zu treffen.

2. **Zeitreihendaten:** GenAI-Modelle, insbesondere Transformer-Modelle, bieten Unternehmen die Möglichkeit, komplexe Muster in Zeitreihendaten zu erkennen und präzise Vorhersagen zu treffen. Dies ermöglicht eine fundierte Entscheidungsfindung und Prozessoptimierung und führt zu Wettbewerbsvorteilen und Effizienzsteigerungen.

3. **Graphen:** GenAI-Modelle analysieren Beziehungen in Graphen und interpretieren komplexe Netzwerke, was kontextbasierte Analysen ermöglicht. Dies bietet Unternehmen tiefere Einblicke in ihre Daten und unterstützt strategische

Entscheidungen, die zu einer besseren Vernetzung und Optimierung von Geschäftsprozessen führen.

4. **Digitale Zwillinge:** Spezialisierte Modelle integrieren relevante Datenquellen und ermöglichen intuitive Interaktionen mit Digitalen Zwillingen. Indem sie die reale Welt präzise abbilden, verbessern sie die Effizienz und Genauigkeit von Simulationen und unterstützen die Optimierung von Geschäftsprozessen.

5. **Multimodale Daten:** In der Robotik integrieren multimodale Daten verschiedene Quellen, um kontextsensitive Entscheidungen zu treffen. Dies erhöht die Flexibilität und Effizienz autonomer Systeme, verbessert die Mensch-Maschine-Interaktion und sorgt für innovative Lösungen und Wettbewerbsvorteile.

6.2. Die Zukunft in Unternehmen mit GenAI

In Anbetracht der rasend fortschreitenden Entwicklungen im Bereich der GenAI zeichnen sich vielversprechende Perspektiven für Unternehmen ab. Die Forschung steht an der Schwelle zur nächsten großen Innovation: KI-Agenten. Diese fortschrittlichen Systeme gehen weit über die Rolle intelligenter Assistenten hinaus und markieren den Beginn einer neuen Ära der Automatisierung. Derzeit liegt der Fokus auf GenAI für intelligente Assistenten, die durch Integration in Unternehmensprozesse und -systeme Personal unterstützen. Doch mit zunehmendem Vertrauen bietet es sich an, solchen Assistenten mehr Befugnisse und Autonomie zu geben und sie schrittweise zu autonomen Agenten weiter zu entwickeln.

KI-Agenten »orchestrieren« nicht nur ein GenAI-Modell, sondern kombinieren mehrere, um ihre Aufgaben effizient zu erfüllen. Sie nutzen beispielsweise ein Modell für multimodalen Input, ein Reasoning-Modell für die Planung und Action-Modelle für die Bedienung von Browsern, anderen digitalen Schnittstellen, Software, Robotern sowie Maschinen und Geräten. Diese Agenten sind flexibel und können den Prozess jederzeit unterbrechen, um Nutzerinnen und Nutzer einzubeziehen, Sicherheitsprüfungen vorzunehmen oder auf konventionelle Software zurückzugreifen. Zudem sind sie in der Lage, den gesamten Prozess zu dokumentieren und aus ihren eigenen Erfahrungen zu lernen, was ihre Effizienz und Anpassungsfähigkeit kontinuierlich verbessert.

Inzwischen sind mit AutoGen, CrewAI, LangGraph, OpenAI Agents SDK, Google Agent Development Kit schon eine ganze Reihe von Frameworks für Multi-Agenten-Systeme entstanden. Hinzu kommt das Model Context-Protokoll (MCP), das den Zugriff von Agenten auf externe Systeme standardisiert, und das Agent-to-Agent-Protokoll (A2A) für die Interaktion zwischen Agenten verschiedener Hersteller. All diese Software-Technik professionalisiert und skaliert die Entwicklung von KI-Agenten mit plattformübergreifenden Befugnissen.

Durch die Integration solcher KI-Agenten können Unternehmen nicht nur ihre Automatisierungsprozesse optimieren, sondern auch neue Geschäftsmöglichkeiten erschließen. Sie heben die Automatisierung auf eine neue Stufe, indem sie komplexe Aufgaben mit höherer Präzision und Geschwindigkeit bewältigen und gleichzeitig die Interaktion mit bestehenden Systemen und Prozessen verbessern. Für Unternehmen bedeutet dies eine zusätzliche und erhöhte transformative Chance, Wettbewerbsvorteile durch innovative und adaptive Technologien zu sichern.

Während generell die Chancen von GenAI für Unternehmen enorm sind, gilt es bei der Einführung, die Herausforderungen frühzeitig zu bedenken und zu antizipieren:

- **Datenqualität:** Die Qualität der Daten muss im Vorfeld sichergestellt werden – denn unvollständige, fehlerhafte oder widersprüchliche Eingabedaten beeinträchtigen die Leistungsfähigkeit und Zuverlässigkeit der Ausgaben erheblich. Eine effektive Data Governance ist daher unerlässlich, um die Qualität, Sicherheit und Integrität der Daten zu gewährleisten. Sie umfasst Richtlinien und Verfahren zur Verwaltung von Daten über ihren gesamten Lebenszyklus hinweg und stellt sicher, dass Daten konsistent und zuverlässig sind. Daher ist es entscheidend, gezielt in die Erhebung und Aufbereitung von Daten zu investieren, um die Grundlage für erfolgreiche GenAI-Anwendungen zu schaffen. Die Chance: GenAI-Modelle können auch die Qualität der Daten, prüfen, sie verbessern und ergänzen oder zusätzliche synthetische Daten erzeugen.

- **Datensicherheit und Souveränität:** Der Schutz sensibler Unternehmens- und Kundendaten sowie die Kontrolle über ihre Verwendung stellt Unternehmen vor Herausforderungen. Eine starke Data Governance hilft dabei, die Nutzung und den Schutz dieser Daten zu optimieren und gleichzeitig regulatorische Anforderungen zu erfüllen. In diesem Kontext bieten Open-Source-Modelle eine flexible Lösung, da sie in KI-Anwendungen lokal oder innerhalb nationaler Grenzen betrieben werden können, wodurch die Datenverarbeitung besser kontrolliert werden kann. Mit Open-Source-Modellen können Unternehmen die Technologie an ihre Sicherheitsanforderungen anpassen.

- **Aktualisierung der Modelle:** Unternehmen müssen sicherstellen, dass ihre KI-Anwendungen mit aktuellen und korrekten Daten arbeiten, um falsche Schlussfolgerungen zu verhindern. Die Aktualisierung von Modellen ist aber ressourcenintensiv. Sinnvoll kann hier die Investition in kleinere, spezialisierte Modelle oder Retrieval-Augmented Generation (RAG)-Architekturen sein, die mit wenig Aufwand tagesaktuell gehalten werden können.

- **Technisches Verständnis:** Für die Entwicklung und Bewertung von Use Cases ist ein fundiertes Verständnis der technischen Möglichkeiten und aktuellen Fortschritte von GenAI unerlässlich. Unternehmen können die Effektivität ihrer Teams durch eine kompetente Begleitung verbessern.

- **Wirtschaftlichkeit der Umsetzung:** Die wirtschaftliche Rentabilität spielt eine entscheidende Rolle. Unternehmen sollten genau prüfen, welche KI-Technologien die gewünschten Ergebnisse liefern und sich langfristig auszahlen, und dann gezielt investieren. Build-or-Buy bedeutet bei GenAI oft die Entscheidung zwischen einem geschlossenen oder offenen Modell, und eine Abwägung der Anpassungs- und Erweiterungsmöglichkeiten. GenAI-Frameworks erleichtern den Wechsel zwischen Modellen, und Protokolle standardisieren den Zugriff auf externe Schnittstellen. In einem sich schnell ändernden Umfeld sollte man sich möglichst wenig an einen Anbieter binden.

- **Befähigung der Mitarbeitenden:** Um mit KI einen Mehrwert zu schaffen, braucht es alle Mitarbeitenden: KI sollte im Alltag mitgedacht werden, um neue Einsatzmöglichkeiten zu erkennen. Eine kontinuierliche Weiterbildung stellt sicher, dass die Beschäftigten die Technologie einschätzen und effektiv einsetzen können. Eine Weiterbildung umfasst also nicht nur Anwendungswissen, sondern auch die Entwicklung eines strategischen Verständnisses für die Integration von GenAI in bestehende Geschäftsprozesse.

- **Fachkräftemangel:** Der Mangel an qualifizierten Fachkräften ist ein ernstzunehmendes Problem, das viele Unternehmen daran hindert, ihre KI-Projekte voranzutreiben. Unternehmen sollten Strategien entwickeln, um diese Talentlücke zu schließen und Mitarbeitende zu gewinnen oder weiterzubilden. Gleichzeitig kann GenAI dazu beitragen, die Arbeitslast des vorhandenen Personals zu reduzieren, indem es Routineaufgaben übernimmt und so den Mitarbeitenden ermöglicht, sich auf strategisch wichtigere Tätigkeiten zu konzentrieren.

- **Erfüllung regulatorischer Anforderungen:** Im digitalen Umfeld entstehen immer mehr Gesetze. Insbesondere in Regionen mit strengen gesetzlichen Regelungen kommen Unternehmen an den Themen Datensicherheit und Datenschutz nicht vorbei. Eine proaktive Auseinandersetzung mit dem EU AI Act und ein verantwortungsbewusster Umgang mit KI-Technologien sowie die Implementierung ethischer Richtlinien lohnt sich und stärkt darüber hinaus das Vertrauen der Mitarbeiterinnen und Konsumenten.

- **Einbettung in Unternehmensstrukturen:** Die Einbettung von KI in bestehende Unternehmensstrukturen kann sich als komplex und zeitintensiv erweisen. Um eine Balance zwischen dezentraler Freiheit für Innovationen und zentraler Kontrolle zur Sicherstellung der Governance zu finden, hilft es, die Unternehmenskultur auf den Prüfstand zu stellen. Um ein Unternehmen an die neuen Gegebenheiten anzupassen, müssen die vorhandenen Prozesse geprüft und erneuert werden.

6.3. GenAI als Gamechanger – Empfehlungen für die Erschließung komplexer Unternehmensdaten

Angesichts der Dynamik von GenAI und sich ständig erweiternden Anwendungsperspektiven empfiehlt sich dennoch ein schrittweiser und umsichtiger Ansatz, um die Komplexität zu beherrschen und nachhaltige Lösungen zu entwickeln.

Unternehmen sollten sich zunächst grundsätzlich mit dem Thema GenAI auseinandersetzen, Potenziale und Anwendungsmöglichkeiten im Unternehmen identifizieren und eine Einsatzstrategie entwickeln. Es ist sinnvoll, dies in Begleitung von erfahrenen anwendungsorientierten Fachleuten durchzuführen. So können potenzielle Anwendungsmöglichkeiten direkt auf Umsetzbarkeit und Wirtschaftlichkeit hin diskutiert und geprüft werden.

Die Allianz Big Data und Künstliche Intelligenz der Fraunhofer-Gesellschaft mit über 30 interdisziplinär arbeitenden Instituten bietet wertvolle Expertise und Unterstützung im Bereich Künstliche Intelligenz. Sie verfügt über den Zugang zu Fachwissen aus verschiedensten Branchen und unterstützt Unternehmen dabei, Strategien zur digitalen Transformation zu entwickeln und umzusetzen. Das technologische Angebot der Allianz bietet umfassende Unterstützung und innovative Werkzeuge, die von der Orientierung bis zur Absicherung implementierter GenAI-Modelle reichen.

Um GenAI als Gamechanger im Unternehmen zu etablieren, sollten Unternehmen folgende Schritte beachten. Die Unterstützung durch erfahrene Partner kann bei allen Schritten von großem Vorteil sein:

1. **Orientierung & Bedarf:**
   - **Grundlagen verstehen:** Beginnen Sie mit einer umfassenden Einführung in die GenAI, um deren Funktionsweise und Potenziale zu verstehen.
   - **Potenzial analysieren:** Identifizieren Sie spezifische Geschäftsbereiche, die von GenAI profitieren könnten. Führen Sie Workshops und Schulungen durch, um das Wissen im Unternehmen zu verbreiten.

2. **Strategie & Konzept:**
   - **Ziele definieren:** Setzen Sie klare Ziele für die GenAI-Implementierung, die mit der Unternehmensstrategie übereinstimmen.

	**Inspire**	Orientierung & Bedarf	Erste Einblicke, Potenziale und Anwendungsideen entdecken
	**Design**	Strategie & Konzept	Strategische Konzeptentwicklung einer KI Lösung
	**Develop & Apply**	Entwicklung & Integration	Entwicklung & Integration von maßgeschneiderten Lösungen
	**Responsibly**	Sicher & Nachhaltig	Nachhaltige Verankerung der Lösung

Abbildung 35: Das Leistungsangebot der Fraunhofer-Allianz Big Data und Künstliche Intelligenz
Quelle: Fraunhofer IAIS

- **Use Cases identifizieren:** Identifizieren und entwickeln Sie konkrete Anwendungsfälle aus Bereichen in ihrem Unternehmen, die den größten Mehrwert bieten. Berücksichtigen Sie dabei sowohl interne Prozesse als auch Kundeninteraktionen.
- **Technologieauswahl:** Evaluieren Sie verschiedene KI-Modelle und Plattformen, um gezielt herauszufinden, welche am besten zu den spezifischen Anforderungen und Zielen Ihres Unternehmens passen könnten. Es geht darum, die Technologie auszuprobieren und erste Erfahrungen zu sammeln.

3. Entwicklung & Integration:

- **Pilotprojekte starten:** Beginnen Sie mit kleinen, kontrollierten Projekten, um die Machbarkeit und den Nutzen von GenAI-Methoden in ihrem Unternehmen zu testen.
- **Technologieauswahl:** Wählen Sie eine passende Technologie aus, die den Anforderungen des Unternehmens entspricht.
- **Modellanpassung:** Passen Sie das GenAI-Modell an, indem Sie es nachtrainieren, Prompting-Techniken anwenden oder RAGs nutzen, um verschiedene Anwendungsfälle abzudecken.
- **Iterative Entwicklung:** Verwenden Sie Feedback aus den Pilotprojekten, um die Lösungen kontinuierlich zu verbessern und anzupassen.
- **Integrationsplanung:** Entwickeln Sie einen detaillierten Plan, um GenAI-Lösungen in bestehende Systeme und Prozesse zu integrieren.

4. Sicher & Nachhaltig:

- **Datenschutz und Sicherheit:** Implementieren Sie Sicherheitsmaßnahmen und Datenschutzrichtlinien, um die Datenintegrität zu gewährleisten.
- **Risikomanagement:** Entwickeln Sie Strategien zur Risikominderung und sichern Sie Ihre Anwendungen.
- **Mitarbeitendenschulung:** Schulen Sie Ihre Mitarbeitenden kontinuierlich, um den effektiven Einsatz der KI-Lösungen sicherzustellen.
- **Langfristige Nachhaltigkeit:** Überwachen und verbessern Sie die GenAI-Anwendungen kontinuierlich, messen Sie den Erfolg mit KPIs und holen Sie regelmäßig Feedback ein.

Die Integration von GenAI in Geschäftsprozesse sollte als strategische Investition in die Zukunft betrachtet werden, um sich an dynamische Rahmenbedingungen anzupassen, zukünftige Trends zu gestalten und Wettbewerbsvorteile zu erlangen. Eine bloße Reaktion auf aktuelle Entwicklungen greift zu kurz.

Entscheidend ist die Nutzung spezifischer Lösungen zur effizienten Verarbeitung und Interpretation komplexer Unternehmensdaten. Unternehmen, die sich proaktiv mit diesen Technologien auseinandersetzen, werden von GenAI als Gamechanger profitieren. Die Fraunhofer-Allianz Big Data und Künstliche Intelligenz unterstützt dabei, innovative Lösungen zu entwickeln und sich optimal im Markt zu positionieren.

Referenzen

Einleitung

[1] Kintz, M., et al. (2024): *Potenziale Generativer KI für den Mittelstand – Wie große KI-Modelle die Arbeitswelt verändern.* Fraunhofer IAO. Verfügbar unter: https://publica.fraunhofer.de/entities/publication/ae3489ac-b3a5-4316-8218-3d6bbc9343a9/details [zuletzt aufgerufen: 07.07.2025].

[2] Paaß, G., & Hecker, D. (2020): *Künstliche Intelligenz: Was steckt hinter der Technologie der Zukunft?* Springer Vieweg.

[3] Vaswani, A., et al. (2017): *Attention is all you need.* NeuRIPS 2017.

[4] Radford, A., et al. (2018): *Improving Language Understanding by Generative Pre-Training.* OpenAI.

[5] Paaß G., & Hecker, D. (2024): *Artificial Intelligence – What is behind the Technology of the Future?* Springer Nature. Verfügbar unter: https://link.springer.com/book/10.1007/978-3-031-50605-5. [zuletzt aufgerufen: 07.07.2025].

[6] Hecker, D., et al. (2023): *Big Data 2.0 – mit synthetischen Daten KI-Systeme stärken.* Wirtschaftsinformatik & Management, 15(2), 161–167.

[7] Paaß, G., & Giesselbach, S. (2023): *Foundation Models for Natural Language Processing: Pre-trained Language Models Integrating Media.* Springer-Nature. Verfügbar unter: https://link.springer.com/10.1007/978-3-031-23190-2 [zuletzt aufgerufen: 07.07.2025].

Tabellen und Matrizen

[8] Fang, X., et al. (2024): *Large Language Models (LLMs) on Tabular Data: Prediction, Generation, and Understanding – A Survey.* arXiv preprint arXiv:2402.17944, Amazon Science. Verfügbar unter: https://assets.amazon.science/f1/0c/5f8a587c452d9b1c687f70b731ab/large-language-models-llms-on-tabular-data-prediction-generation-and-understanding-a-survey.pdf [zuletzt aufgerufen: 04.06.2025].

[9] Ma, J., et al. (2024): *Tabdpt: Scaling tabular foundation models.* arXiv preprint arXiv:2410.18164, Verfügbar unter: https://arxiv.org/pdf/2410.18164 [zuletzt aufgerufen: 04.06.2025].

[10] Kang, et al. (2025): *Tabular transformer generative adversarial network for heterogeneous distribution in healthcare.* Scientific Reports, 15(1), 10254. Verfügbar unter: https://www.nature.com/articles/s41598-025-93077-3#:~:text=information,dataset%2C%20the%20discretization%20and%20converter [zuletzt aufgerufen: 04.06.2025].

[11] Hollmann, N., et al. (2025): *Accurate predictions on small data with a tabular foundation model.* Nature, 637(8045), 319-326. Verfügbar unter: https://www.nature.com/articles/s41586-024-08328-6 [zuletzt aufgerufen: 04.06.2025].

[12] Chen, T., & Guestrin, C. (2016): *Xgboost: A scalable tree boosting system.* In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August, pp. 785-794. Verfügbar unter: https://arxiv.org/abs/1603.02754 [zuletzt aufgerufen: 04.06.2025].

[13] Gardner, J., et al. (2024): *Large scale transfer learning for tabular data via language modeling.* arXiv preprint arXiv:2406.12031, arXiv.org.

Zeitreihen

[14] Business Research Insights (2025): *Time Series Forecasting Market Size, Share, Growth, And Industry Analysis, By Type (Software, Service), By Application (Business Planning, Financial Industry, Medical), Regional Insights and Forecast to 2033.* Business Research Insigshts. Verfügbar unter: https://www.businessresearchinsights.com/market-reports/time-series-forecasting-market-114943. [zuletzt aufgerufen: 03.07.2025].

[15] Burr, J., et al. (2023): *Plastic extrusion process optimization by inversion of stacked autoencoder classification machines.* Chemie Ingenieur Technik 95.10 : 1555–1562.

[16] Kourani, H., et al. (2024): *ProMoAI: Process Modeling with GenAI.* International Joint Conferences on Artifical Intelligence Organization. Verfügbar unter: https://www.ijcai.org/proceedings/2024/1014 [zuletzt aufgerufen: 07.07.2025].

[17] Kourani, H., et al (2024): *Leveraging Large Language Models for Enhanced Process Model Comprehension.* SSRN. Verfügbar unter:https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5017379 [zuletzt aufgerufen: 07.07.2025].

[18] Garza, A., et al. (2023): *TimeGPT-1.* Cornell University. Verfügbar unter: https://arxiv.org/abs/2310.03589 [zuletzt aufgerufen: 07.07.2025].

[19] Rasul, K., et al. (2024): *Lag-Llama: Towards Foundation Models for Probalistic Time Series Forecasting.* Verfügbar unter https://arxiv.org/abs/2310.08278 [zuletzt aufgerufen: 07.07.2025].

[20] Liu, Y., et al. (2024): *Timer: Generative Pre-trained Transformers Are Large Time Series Models.* Cornell University. Verfügbar unter: https://arxiv.org/abs/2402.02368 [zuletzt aufgerufen: 07.07.2025].

[21] Meyer, M., et al. (2024): *Benchmarking Time Series Foundation Models for Short-Term Household Electricity Load Forecasting.* Cornell University. Verfügbar unter: https://arxiv.org/pdf/2410.09487v1 [zuletzt aufgerufen: 07.07.2025].

[22] Fraunhofer IOSB-AST: *Energie Einfach machen.* Verfügbar unter: https://www.edm-prophet.de/[zuletzt aufgerufen: 07.07.2025].

[23] Graß, A., (2024): *The Semantic Time Series Ontology.* Verfügbar unter. https://w3id.org/semts [zuletzt aufgerufen: 07.07.2025].

[24] Ott, J., et al. (2024): *Radio Foundation Models: Pre-training Transformers for 5G-based Indoor Localizatio.* Erschienen in: 4th International Conference on Indoor Positioning and Indoor Navigation (IPIN). IEEE. pp. 1–6. Verfügbar unter: 10.1109/IPIN62893.2024.10786154. [zuletzt aufgerufen: 07.07.2025].

[25] Kim, Y., & Moon, W. (2016): *Human detection and activity classification based on micro-Doppler signatures using deep convolutional neural networks.* IEEE Geoscience and Remote Sensing Letters, 13(1), 8–12.

[26] Ma, Y., Geng, et al. (2019): *WiFi sensing with channel state information: A survey.* ACM Computing Surveys (CSUR), 52(3), 1–36. Verfügbar unter: https://dl.acm.org/doi/10.1145/3310194 [zuletzt aufgerufen: 22.05.2025].

[27] Du, H., et al. (2019): *A three-dimensional deep learning framework for human behavior analysis using range-Doppler time points.* IEEE Geoscience and Remote Sensing Letters, 17(4), 611–615.

[28] Hu, E. J., et al (2021): *LoRA: Low-Rank Adaptation of Large Language Models.* arXiv preprint, arXiv:2106.09685. Verfügbar unter: https://arxiv.org/abs/2106.09685 [zuletzt aufgerufen: 22.05.2025].

[29] Gusland, D., et al (2021): *Open Radar Initiative: Large Scale Dataset for Benchmarking of micro-Doppler Recognition Algorithms.* Erschienen in: 2021 IEEE International Radar Conference (RADAR). IEEE.

[30] Stove, A. G. (1992): *Linear FMCW radar techniques.* IEE Proceedings F - Radar and Signal Processing, 139(5), 343–350.

[31] Chen, V. C., et al. (2006): *Micro-Doppler effect in radar: phenomenon, model, and simulation study.* IEEE Transactions on Aerospace and Electronic Systems, 42(1), 2–21. Verfügbar unter: https://ieeexplore.ieee.org/document/1603402 [zuletzt aufgerufen: 22.05,2025].

[32] Tao et al. (2025): *LLM-based framework for bearing fault diagnosis.* Erschienen in: Mechanical Systems and Signal Processing 224.

[33] Purohit et al., (2019): *MIMII Dataset: Sound Dataset for Malfunctioning Industrial Machine Investigation and Inspection.* Erschienen in Proc. 4th Workshop on Detection and Classification of Acoustic Scene and Events (DCASE).

Graphen für komplexe Netzwerke

[34] Li et al. (2024): *A Survey of Graph Meets Large Language Model: Progress and Future Directions.* International Joint Conferences on Artifical Intelligence Organization. Verfügbar unter: https://www.ijcai.org/proceedings/2024/898 [zuletzt aufgerufen: 07.07.2025].

[35] Tang, J., et al. (2024): *GraphGPT: Graph Instruction Tuning for Large Language Models*. Association for Computing Machinerey. Verfügbar unter: https://dl.acm.org/doi/10.1145/3626772.3657775 [zuletzt aufgerufen: 07.07.2025].

[36] Yuan, C., et al. (2025): *A Survey of Graph Transformers: Architectures, Theories and Applications.* Cornell University. Verfügbar unter: http://arxiv.org/abs/2502.16533.

[37] Fraunhofer IOSB: *KI in LMS.* Verfügbar unter: https://kilms.fraunhofer.de/ [zuletzt aufgerufen: 07.07.2025].

[38] Fraunhofer SCAI: *Sonar.* Verfügbar unter: www.sonar-redox.eu [zuletzt aufgerufen: 07.07.2025].

[39] Dobberstein, N,. et al. (2024): *Llamol: a dynamic multi-conditional generative transformer for de novo molecular design.* Journal of Cheminformatics. Verfügbar unter: https://github.com/Fraunhofer-SCAI/llamol [zuletzt aufgerufen: 07.07.2025].

[40] Oh, Y., et al. (2022): *Time delay estimation of traffic congestion propagation due to accidents based on statistical causality.* Electronic Research Archive 31(2):691 – 707. Verfügbar unter: https://www.researchgate.net/figure/Locations-of-traffic-accidents-and-their-significant-propagation-paths-in-the-city-of_fig2_365678080 (Copyright: https://creativecommons.org/licenses/by/4.0/)] [zuletzt aufgerufen: 07.07.2025].

Digitale Zwillinge

[41] Emmelmann, M., et al. (2022): *Open 5G campus networks: key drivers for 6G innovations.* Springer Nature Link. Volume 139:589-600. Verfügbar unter: https://link.springer.com/article/10.1007/s00502-022-01064-7 [zuletzt aufgerufen:07.07.2025].

[42] Fraunhofer IOSB (2025): *Advanced Occupant Monitoring System – Neues Whitepaper zur Aktivitätserfassung im Fahrzeuginnenraum veröffentlicht.* Verfügbar unter: https://www.iosb.fraunhofer.de/de/projekte-produkte/advanced-occupant-monitoring-system.html [zuletzt aufgerufen: 07.07.2025].

Autorinnen und Autoren

Einleitung, Ausblick sowie redaktionelle Leitung

Prof. Dr. Dirk Hecker, Dr. Sonja Holl-Supra, Miriam Mock, Dr. Gerhard Paass, Dr. Angelika Voß

Matrizen und Tabellen

	Use Case
Pascal Nicolai Linden Dr. Tim Wirtz	Voraussage von strukturierten Geschäftsdaten
Dr. Daniel Scherer Dr. Christian Ufrecht	Synthetisierte Quantenschaltungen

Zeitreihen

	Use Case
Julia Burr	Assistenzsystem zur Parameterwahl für Produktionsprozesse
Franziska Dechert Bjarne Huuck Prof. Dr. Wolfgang Kratsch Dr. Gyunam Park Prof. Dr. Maximilian Röglinger Niklas Uhl Prof. Dr. Wil van der Aalst	Eventdaten als Treiber für moderne Process Intelligence
Dr. André Kummerow Florian Rippstein	Large Timeseries Modelle zur Prognose von Energiezeitreihen
Dr. Christian Beecks Alexander Graß	Agentengestützte interaktive Datenanalyse
Dr. Christopher Mutschler Dr. Felix Ott Jonathan Ott Dr. Maximilian Stahlke	Pretrained Transformers für die funkbasierte Lokalisierung
Alper Yaman	Radarbasierte Objekterkennung und -klassifikation
Hartmut Eigenbrod Julian Raible Christian Seidler	Sparsame Zustandsüberwachungen in industriellen Maschinenparks

Graphen für komplexe Netzwerke

	Use Case
Dr. Gesa Benndorf	Sprachmodelle für die Anlagensteuerung und -überwachung
Dr. Christopher Krauss	Lernassistent für die Personalisierung in Lernplattformen
Dr. Jan Hamaekers	De Novo Molekulardesign für gezielt ausgelegte Substanzen
Dr. Georg Fuchs Daniel Schulz	Prognose von Verkehrsmengen für die Standortplanung
Matthis Leicht	Eine flexible Datenstruktur für effiziente Schiffsüberwachung
Varun Gowtham	Flexibles Management für Telekommunikationsnetzwerke

Digitale Zwillinge

	Use Case
Dr. Christian Kühnert Dr. Andreas Wilde Dr. Mathias Ziebarth	Digitaler Zwillinge für effiziente Gebäudebewirtschaftung
Dr. Frederik Diederichs	Digitaler Zwilling von Fahrzeuginsassen
Dr. Ina Steinmetz Dr. Ljiljana Stojanovic	Der digitale Produktpass
Wilhelm Jaspers Dr. Jonathan Krauß	Semantisches Datenmapping bei digitalen Zwillingen

Multimodale Daten und Robotik

	Use Case
Dr.-Ing. Janko Petereit	Steuerung von autonomen Baumaschinen
Alexander Zorn	Robotik & Anlagenprozesse mit Sprach- & KI-Agenten steuern
Nicolas Bach Maximilian Hörstrup	Revolutionierung der Kommissionierung: KI für die Logistik
Christian Pionzewski Andrej Plenne	Intelligente Produktbeschreibung bei Retouren und Re-Commerce
Dr. Florian Strohm	Kollaborative Robotik mit Pose- und Gaze-Tracking

Abbildungsverzeichnis

Abbildung 1: Schema eines GPT-Models 7

Abbildung 2: Der Transformer-Block besteht aus einem »Self-Attention-Modul«. 8

Abbildung 3: Ein Foundation-Modell erhält eine Eingabesequenz aus einer oder mehreren Medienarten. 1

Bildnachweise: Bild Graph: Wikipedia - My Friend (CC BY-SA 3.0), https://commons.wikimedia.org/wiki/File:U-Bahn-Netz_Wien_2019.png [zuletzt aufgerufen: 22.07.2025], Bild Digitaler Zwilling: Wikipedia - SumitAwinash (CC BY-SA 4.0), https://commons.wikimedia.org/wiki/File:Oil_rig_Jan_23.jpg [zuletzt aufgerufen: 22.07.2025], Bild Roboterdaten: Wikipedia - NASA/JPL/Cornell University, Maas Digital LLC (Public Domain), https://commons.wikimedia.org/wiki/File:NASA_Mars_Rover.jpg [zuletzt aufgerufen: 07.07.2025]

Abbildung 4: Beispiel einer Tabelle und einer Matrix 12

Abbildung 5: GenAI mit Tabellen & Matrizen 13

Abbildung 6: Vereinfachte Darstellung der Voraussage mit LLMs 15

Abbildung 7: Zwei Quantenschaltungen, bei denen ein RL-Algorithmus die ursprüngliche Schaltung als unitäre Matrix optimiert, um die Synthese ohne T-Gatter zu erreichen. 16

Abbildung 8: Die Datenstruktur Zeitreihen 17

Abbildung 9: Strukturierter Überblick, wie Zeitreihen in den nachfolgenden Use Cases verarbeitet und GenAI eingesetzt werden kann. 18

Abbildung 10: Die vorgegebenen Prozessparameter führen zum gewünschten Temperaturverlauf der Schmelze. 20

Abbildung 11: Referenzarchitektur von GenAI in Process Intelligence 21

Abbildung 12: Vergleich Fachmodelle mit LTMs zur Vorhersage von Energiezeitreihen 22

Abbildung 13: Interface der aktuellen Version des agentengestützen Datenanalysetools SAGE 23

Abbildung 14: Unüberwachter Pre-Training-Prozess: Die Aufgabe des Modells ist es, vorher entfernte Signalteile zu rekonstruieren. 24

Abbildung 15: Doppler-Frequenzverschiebungen vs. Laufzeit für Radfahrer, UAV, Fußgänger, und Fahrzeuge. 25

Abbildung 16: Konzept zum Einsatz von LLMs (nach Tao [32]) 26

Abbildung 17: Ein Knotensystem dargestellt als Graph. 27

Bildnachweis: Wikipedia Bernese media, https://de.wikipedia.org/wiki/Datei:Knotensystem_Schweiz_mit_Bahn2000_2.Etappe_2030.png, CC BY-SA 3.0 [zuletzt aufgerufen: 22.07.2025]

Abbildung 18: Verschiedene Repräsentationen der Graphen, die in unterschiedliche Typen von Embeddings transformiert werden. . . . 28

Abbildung 19: Ausschnitt aus einem Stromlaufplan . . . 30

Abbildung 20: Abbildung der persönlichen Lernstandsvorhersage . . . 31

Abbildung 21: Schematische Darstellung einer Redox-Flow-Batterie einschließlich verschiedener Skalen und des chemischen Suchraums für Elektrolyte . . . 32

Abbildung 22: Ein Graph kann das Straßennetz einer Großstadt beschreiben. . . . 33

Abbildung 23: Räumlich-zeitlicher Graph für Schiffsbewegungen . . . 34

Abbildung 24: KG-GPT Funktionsdiagramm . . . 35

Abbildung 25: Ausprägungsformen von Digitalen Zwillingen . . . 37

Abbildung 26: 3D-Visualisierung eines Gebäudes als Digitaler Zwilling . . . 39

Abbildung 27: Computer Vision Analyse der Fahrzeuginsassen . . . 40

Abbildung 28: Das FA³ST Ecosystem vereint das Domänenwissen und die Daten der Nutzenden mit der VWS-Expertise und LLM-Fähigkeiten. . . . 41

Abbildung 29: Semantisches Datenmapping in der Umsetzung . . . 42

Abbildung 30: Autonomer Bagger des Fraunhofer IOSB beim Beladen eines Brechers . . . 45

Abbildung 31: Fraunhofer IAIS Demonstrator – ein per Sprache programmierbarer Roboterarm . . . 46

Abbildung 32: Pose und Blickrichtung des Menschen wird erkannt während der Kollaboration mit einem Roboter. . . . 47

Abbildung 33: Mobiler Manipulatoren-Prototyp des Fraunhofer IML: mögliche Plattform für den Kommissionierungs-Agenten . . . 48

Abbildung 34: Scan einer Retoure zum Versand . . . 49

Abbildung 35: Das Leistungsangebot der Fraunhofer-Allianz Big Data und Künstliche Intelligenz . . . 53

Impressum

Kontakt:
Fraunhofer-Allianz Big Data und Künstliche Intelligenz
Schloss Birlinghoven 1
53757 Sankt Augustin
www.bigdata-ai.fraunhofer.de

Herausgeber:
Fraunhofer-Allianz Big Data und Künstliche Intelligenz

Redaktion:
Dr. Sonja Holl-Supra, Fraunhofer BDAI
Miriam Mock, Fraunhofer BDAI

Layout und Satz:
Achim Kapusta, Fraunhofer IAIS
Lea Scharmach Silva, Fraunhofer IAIS

Grafik:
Achim Kapusta, Fraunhofer IAIS
Asra Soraya Neumeister, Fraunhofer IAIS
Lea Scharmach Silva, Fraunhofer IAIS (Titelbild)

Titelbild:
Titelseite KI-generiert mit Copilot

Bibliografische Information der Deutschen Nationalbibliothek:
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.de abrufbar.

ISBN (Printausgabe): 978-3-8396-2133-2
DOI (kostenlose Open-Access-Version):
https://doi.org/10.24406/publica-4938

Druck und Weiterverarbeitung:
Schleunungdruck GmbH, Marktheidenfeld

Für den Druck des Buches wurde chlor- und säurefreies Papier verwendet.



Dieses Werk ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz:
https://creativecommons.org/licenses/by/4.0/legalcode.de

© Fraunhofer Verlag, 2025
Nobelstraße 12
70569 Stuttgart
verlag@fraunhofer.de
www.verlag.fraunhofer.de

als rechtlich nicht selbständige Einheit der

Fraunhofer-Gesellschaft zur Förderung
der angewandten Forschung e.V.
Hansastraße 27 c
80686 München
www.fraunhofer.de

Generative Künstliche Intelligenz (GenAI) hat sich weit über die reine Text- und Bildverarbeitung hinaus zu einem leistungsfähigen Werkzeug für die Analyse, Interpretation und Optimierung unterschiedlicher Unternehmensdaten entwickelt. Doch wie können Unternehmen dieses Potenzial der Technologie für sich nutzen? Und was leistet die angewandte Forschung jenseits der bekannten Einsatzfelder?

In diesem Whitepaper der Fraunhofer-Allianz Big Data und Künstliche Intelligenz wird aufgezeigt, wie Modelle der generativen KI auch komplexe Datenstrukturen – etwa Tabellen, Zeitreihen, Graphen und andere Formate wie Digitale Zwillinge oder multimodale Sensordaten für Roboter – erschließen können und wie sie auf diese Weise neue Chancen und Mehrwerte für verschiedene Branchen schaffen. Mit praxisnahen Beispielen und wissenschaftlich fundierten Erkenntnissen von Fraunhofer-Forscherinnen und -Forschern bietet dieses Whitepaper einen kompakten und inspirierenden Einblick in das Zukunftsfeld »GenAI« und zeigt auf, welche Perspektiven dieses für Ihr Unternehmen eröffnet.

Fraunhofer VERLAG

ISBN 978-3-8396-2133-2

9 783839 621332