Operatorassistenz für das Scherschneiden mittels agentenbasierter Künstlicher Intelligenz

Marcel Görz M.Sc.[(1)], Dr.-Ing. Adrian Schenek[(1)], Dr.-Ing. Kim Rouven Riedmüller[(1)], Univ.-Prof. Dr.-Ing. Dr. h. c. Mathias Liewald MBA[(1)]

[(1)] Institut für Umformtechnik, Universität Stuttgart, Holzgartenstr. 17, 70174 Stuttgart, Deutschland

1. Einleitung

Das Scherschneiden stellt eines der ökonomisch bedeutendsten Fertigungsverfahren der blechbearbeitenden Industrie dar, da nahezu jedes Blechbauteil im Laufe seiner Wertschöpfung mehrfach mittels dieses Verfahrens beschnitten bzw. gelocht wird. Dabei müssen die erzeugten Schnittkanten aufgrund steigender Bauteilanforderungen häufig eine hohe Qualität hinsichtlich der Maßhaltigkeit und Oberflächenbeschaffenheit aufweisen [1]. Die Einhaltung dieser Qualitätsanforderungen wird jedoch maßgeblich durch die Sensibilität des jeweiligen Scherschneidprozesses gegenüber stochastischen Materialschwankungen (z. B. Variationen in der Zugfestigkeit oder Blechdicke) beeinflusst, welche zu instabilen Prozesszuständen und damit einhergehend zu hohen Ausschussraten führen können. Zur Beherrschung dieser chargenspezifischen Materialschwankungen sind daher häufig Prozessanpassungen erforderlich, die heute üblicherweise durch das Bedienpersonal auf Basis von implizitem Fachwissen („Tacit Knowledge“) vorgenommen werden [2]. Infolge des demografischen Wandels steht die Branche jedoch vor einem massiven „Brain-Drain“, also dem Verlust langjähriger und erfahrener Mitarbeiter, was die Dokumentation des heute noch bestehenden Fachwissens mit Blick auf zukünftige Produktionsprozesse dringend erforderlich macht [3]. Das erfahrungsbasierte Wissen dieser Mitarbeiter zur Kompensation von Prozessfehlern ist durch konventionelle Dokumentationssysteme allerdings kaum abbildbar, sodass hierfür neue Methoden entwickelt und angewandt werden müssen.

Vor diesem Hintergrund wird im vorliegenden Beitrag eine agentische Architektur auf Basis von Large Language Models (LLMs) vorgestellt, die als Cognitive Twin fungiert [4]. Im Gegensatz zu klassischen deskriptiven Digitalen Zwillingen ermöglicht dieser Ansatz eine Orchestrierung heterogener Datenquellen, um z.B. physikalisch fundiertes Reasoning [5] in die Herstellung von Bauteilen mittels Scherschneiden zu integrieren. Ziel ist es, dem Operator durch die Verknüpfung von Analytik und explizitem Fachwissen kausale Entscheidungshilfen zur Unterstützung bei erforderlichen Prozessanpassungen bereitzustellen.

2. Stand der Forschung

2.1 Limitierungen großer Sprachmodelle im industriellen Kontext

Die auf der von Vaswani *et al. (*2017) vorgestellten Transformer-Architektur [6] basierenden und durch ChatGPT [7] bekannt gewordenen Large Language Models (LLMs) weisen eine hohe Leistungsfähigkeit in der Sprachverarbeitung auf, jedoch bestehen Hindernisse hinsichtlich deren Einsatz in der Überwachung von Fertigungsprozessen. Ein großes Problem stellen hierbei die sogenannten Halluzinationen dar. [8] Diese Halluzinationen basieren darauf, dass LLMs als probabilistische Systeme operieren und damit lediglich die statistische Wahrscheinlichkeit der nächsten

Verlag Meisenbach GmbH, Geisfelder Straße 14, 96050 Bamberg/Deutschland
http://www.umformtechnik.net/whitepaper/

Wortfolge berechnen, ohne über ein echtes Verständnis für physikalische Zusammenhänge oder technische Fakten zu verfügen. Wenn das Modell auf Informationslücken stößt, füllt es diese mit Mustern, die zwar plausibel klingen, aber in der Realität Fehlentscheidungen (z. B. falsche Einstellparameter) hervorrufen können. Weiterhin werden LLMs durch ihren statischen Wissensstand (Knowledge Cutoff) limitiert, da ein LLM nur Wissen besitzen kann, welches zum Trainingszeitpunkt verfügbar war. Implizites Fachwissen von erfahrenen Operatoren, z.B. die Auslegung und Überwachung von Scherschneidprozessen wird deshalb auch nicht berücksichtigt. Ein weiteres Hindernis für den Einsatz von LLMs zur Unterstützung der Herstellung von Bauteilen durch Scherschneiden stellt zudem die fehlende Erklärbarkeit (Explainability) der Modellentscheidungen dar, da die interne Logik der Gewichtungen, die zu den ausgegebenen Entscheidungen führen, für den Anwender eine „Black Box“ darstellt [9].

2.2 Agentic AI und das Model Context Protocol (MCP)

Die Entwicklung von Agentic LLMs [10] markierte den Übergang von der reinen Wortvorhersage zur autonomen Handlungsfähigkeit. Während Reasoning-Modelle [5] durch interne Chain-of-Thought-Prozesse komplexe logische Probleme lösen, erweitern Agentic LLMs diese Basis um eine ausführende Ebene. Diese umfasst die Werkzeugnutzung (aktiver Aufruf externer Software-Tools wie Analyse-Skripte), ein Gedächtnis zur Konsistenzwahrung über komplexe Aufgaben hinweg sowie Feedbackschleifen, die es dem Modell erlauben, Ergebnisse eigener Handlungen zu bewerten und den Handlungsplan iterativ an die Realität anzupassen. Dadurch verändert sich der Denkprozess in einen iterativen Zyklus aus Planung, Handeln und Beobachtung. Dies ermöglicht es dem Modell, nicht nur über Probleme „nachzudenken“, sondern Ziele in dynamischen Umgebungen autonom zu verfolgen [10]. Dieser Trend stellt einen signifikanten Wandel im Vergleich zu klassichen Chatbot-Architekturen wie ChatGPT dar. Das Model Context Protocol (MCP) [11] nimmt hierbei eine Schlüsselrolle ein und ermöglicht die einfache Kommunikation zwischen agentischen LLMs und unterschiedlichsten Daten und Anwendungen. Das MCP stellt als offener Standard eine strikte Trennung zwischen der logischen Abstraktionsebene (Reasoning) und Datenquellen sicher und vereinfacht somit die Integration in sicherheitskritische industrielle IT-Infrastrukturen (Shopfloor-to-Cloud) erheblich [12].

2.3 Spezifische Prozessüberwachung in der Stanztechnik

In der Stanz- und Umformtechnik werden neben klassisch hüllkurvenbasierten Überwachungssystemen gegenwärtig vorrangig Verfahren der Kraft-Weg-Kurvenanalyse sowie Methoden der Anomaly Detection beispielsweise auf Basis von Autoencodern [13] oder Klassifikations-Algorithmen [14] zur Vorhersage von Verschleiß der Schneidaktivelemente angewendet. Mittels dieser Ansätze des Machine Learning können zwar hohe Genauigkeiten bei der Vorhersage des Verschleißes erzielt werden, diese weisen jedoch ein Defizit bei der Integration von implizitem Expertenwissen auf. Zusammenfassend fehlen im Bereich der Stanztechnik generell Systeme, die dem Operator über die Detektion hinaus eine ganzheitliche Unterstützung bei der Entscheidungsfindung zum Umgang mit auftretenden Prozessabweichungen liefern.

3. Systemarchitektur und Implementierung

Aufgrund der im Stand der Technik beschriebenen Defizite wird in diesem Beitrag eine auf einem modularen, agentenbasierten Ansatz basierende Systemarchitektur (vgl. Abbildung 1) vorgestellt, durch welche Domänenkompetenzen für das Scherschneiden mit modernen Modellen zur Sprachverarbeitung kombiniert werden können, um ein neuartiges System zur Unterstützung von Operatoren zu schaffen. Das System gliedert sich in eine zentrale Orchestrierungseinheit auf Basis eines agentischen LLM-Systems sowie in drei spezialisierte Subsysteme, die über MCP verbunden sind. Jedes dieser Subsysteme ist für einen spezifischen Zweck vorgesehen. Eine genaue Beschreibung der einzelnen Teile des Assistenzsystem erfolgt im folgenden Abschnitt.

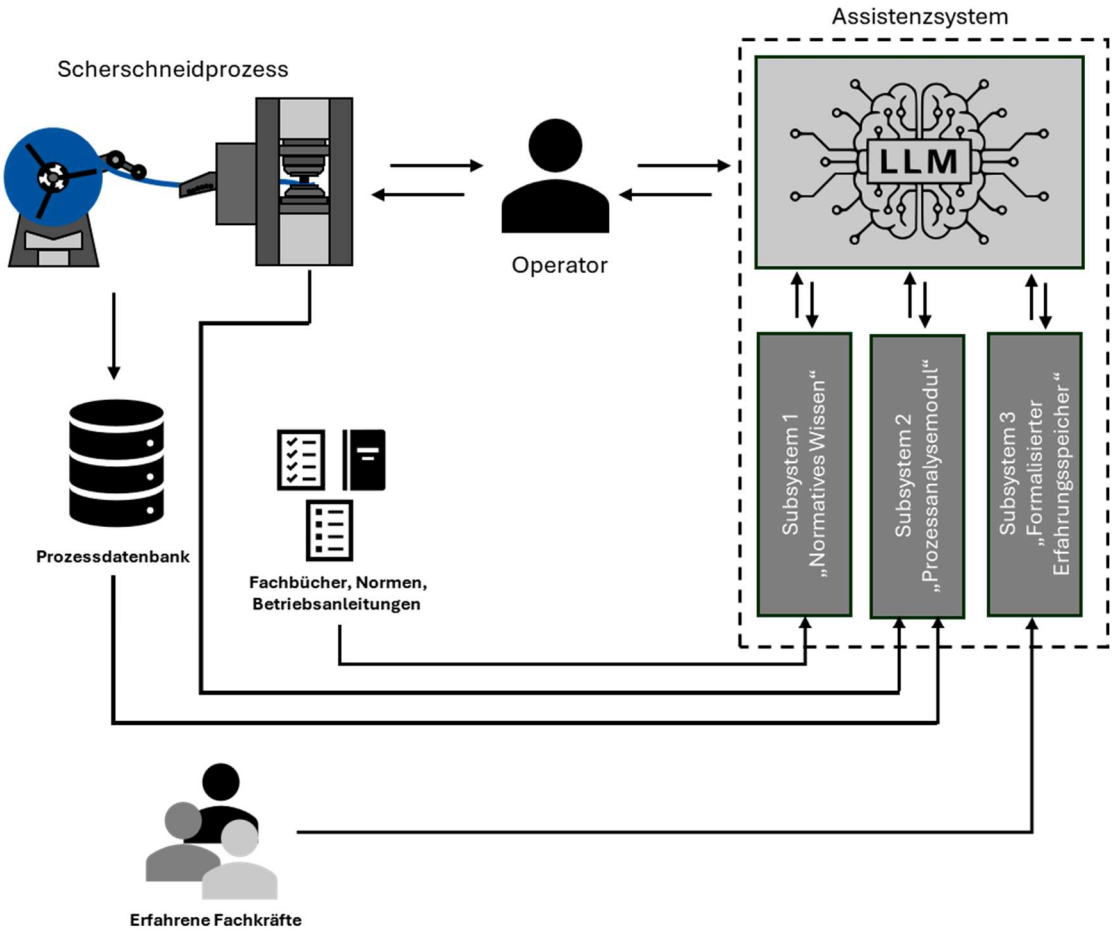

Abbildung 1: Prozessschema für das Scherschneidassistenzsystem

3.1 Der Agent als zentrale Instanz für Scherschneidverfahren

Den Kern des Systems bildet ein agentisches LLM mit hoher Schlussfolgerungskompetenz. Im Gegensatz zu konventionellen Chatbot-Systemen die mit nicht agentischen LLM ausgestattet sind fungiert dieses agentische Modell als autonomer Client. Der Agent erstellt auf Basis der Anfrage durch den Operator einen mehrstufigen Handlungsplan und führt diesen sequenziell aus. Durch die Integration der Subsysteme durch MCP wird sichergestellt, dass das LLM nicht nur reine Textantworten generiert, sondern auf Werkzeuge und Datenbanken zugreift, um zusätzliches Wissen zu nutzen und dadurch korrekte und faktenbasierte Antworten zu generieren.

3.2 Subsystem 1: Normative Wissensbasis

Subsystem 1 fungiert als Wissensbasis und realisiert die Bereitstellung von explizitem Domänenwissen durch ein Retrieval-Augmented Generation (RAG)-System [15] auf Basis einer hochdimensionalen Vektordatenbank. In dieser Datenbank werden explizite Datenquellen wie technische Normen, Maschinendokumentationen und Fachartikel in Form von Einbettungsvektoren (Embeddings) bereitgestellt. Durch den Einsatz von semantischer Suche ist der Agent in der Lage, bei einer Fragestellung relevante Textsegmente in der Datenbank zu finden und diese in seine Antwort zu integrieren. Dieses Vorgehen stellt sicher, dass die Ausgaben des Sprachmodells nicht auf Schätzungen des LLM basieren, sondern auf verifizierten, fachspezifischen Wissensquellen. Dadurch wird die für generative Modelle typische Halluzinationsneigung reduziert, da die Antwort des Agenten auf Basis von Fachliteratur erfolgt.

3.3 Subsystem 2: Prozessanalysemodul

Subsystem 2 fungiert als Schnittstelle zwischen dem Agenten und dem betrachteten Fertigungsprozess. Es ermöglicht eine datengestützte Zustandsbewertung des Fertigungsprozesses basierend auf historische Prozessdaten in einer Datenbank oder auf Messdaten, die unmittelbar mittels in die Stanzpresse oder in das Stanzwerkzeug integrierter Sensorik erfasst werden. Zu den erfassten Zeitreihendaten zählen insbesondere hochfrequente Kraft-Weg-Verläufe oder z. B. thermische Signale aus dem Umformprozesse. Zur Auswertung dieser Daten werden üblicherweise Python-basierte Analysealgorithmen und Software-Werkzeuge eingesetzt, mit welchen Anomalien und Abweichungen im Fertigungsprozess identifiziert werden können.

3.4 Subsystem 3: Formalisierter Erfahrungsspeicher

Subsystem 3 adressiert die Externalisierung von bislang implizitem Erfahrungswissen von Mitarbeitern. Um eine reproduzierbare Entscheidungsfindung zu gewährleisten, wird dieses Expertenwissen in Python-basierte Tools überführt. Diese Formalisierung stellt sicher, dass Strategien zur Fehlerbehebung oder Optimierung von Prozess- und Designparametern regelbasiert durchgeführt werden. Der Agent greift auf diese Säule zurück, um beispielsweise die Optimierung der Geometrie eines Schneidstempels zur Maximierung des Glattschnitts durchzuführen. Solche Optimierungsprozesse werden heute in der Regel durch erfahrene Ingenieure oder Werkzeugmacher im Trial-and-Error-Prinzip und dann schließlich erfahrungsbasiert durchgeführt.

3.5 Einsatzspektrum des Systems

Die modulare Struktur des Assistenzsystems ermöglicht ein duales Einsatzspektrum. Einerseits fungiert das System als strategischer Berater in der Phase der Prozessauslegung. Hierbei unterstützt der Agent Techniker oder Ingenieure bei der konstruktiven Festlegung von Werkzeuggeometrien und Eingabe von Einstellwerten, indem er auf die normative Wissensbasis (Subsystem 1) und Optimierungstools (Subsystem 3) zugreift. Andererseits dient die Architektur der operativen Unterstützung unmittelbar während der Fertigung. In diesem Szenario agiert das System prozessbegleitend, wertet Sensordaten in Echtzeit aus (Subsystem 2) und stellt dem Operator bei auftretenden Instabilitäten sofortige, evidenzbasierte Handlungsempfehlungen zur Verfügung, um Ausschuss weitgehend zu vermeiden.

4. Use Case und Validierung: Optimierung des Glattschnittanteils einer Schnittfläche

Die Leistungsfähigkeit des hier vorgestellten Assistenzsystems wurde am Beispiel des industriellen Anwendungsfalls „Optimierung des Glattschnittanteils einer Schnittfläche“ demonstriert. Der Glattschnittanteil stellt eine wichtige Qualitätskenngröße beim Scherschneiden dar, der bei Unterschreitung eines Mindestwertes zu Ausschuss führt. Tritt während der Fertigung eine Abweichung des Glattschnittes auf, so muss der Schneidprozess derart optimiert werden, dass der Glattschnittanteil wieder über dem definierten Grenzwert liegt. Die technische Realisierung basiert auf dem agentischen LLM Claude 4.5 Sonnet, von Anthropic welches als zentraler Agent fungiert und über das MCP die spezialisierten Software-Werkzeuge aus jedem der drei Subsysteme orchestriert.

Das erste Werkzeug ermöglicht über eine RAG-Architektur den Zugriff auf wissenschaftliche Publikationen zu unterschiedlichen Scherschneidverfahren, woraus der Agent theoretische Einflussfaktoren und verfahrensspezifische Besonderheiten ableiten kann. Parallel dazu dient ein zweites Python-basiertes Werkzeug der quantitativen Prozessanalyse. Dieses extrahiert aus den Kraft-Weg-Verläufen der Presse den realen Glattschnittanteil und stellt diesen als Ist-Zustand bereit. Das zugrunde liegende Machine Learning Modell basiert auf dem von Schenek et al. erstellten Neuronalen Netz und wurde um die notwendigen Befehle zur Abfrage per MCP erweitert [16]. Zur finalen Problemlösung adressiert der Agent schließlich ein Software-Werkzeug zur Stempelgeometrieoptimierung [17], welches ebenfalls für die Abfrage per MCP erweitert wurde. Dieses Software-Werkzeug berechnet auf Basis der vorhandenen Prozessparameter und des gewünschten Glattschnittanteils das notwendige Design des Schneidstempels.

Nachfolgend ist der Prozessablauf innerhalb des Assistenzsystems aufgeführt:

- **Detektion:** Der Werker identifiziert während der Fertigung von Bauteilen eine Abweichung des Glattschnittanteils von den Qualitätsvorgaben und befragt den Assistenten.
- **Analyse:** Das System bestimmt über eine Schnittstelle zur im Umformwerkzeug verbauten Kraftmessung die Glattschnittwerte aus den Produktionsdaten. Im Vergleich zu hinterlegten Sollwerten erkennt er einen Abfall des Glattschnittanteils und damit Handlungsbedarf.
- **Methodenevaluation:** Der Agent analysiert mit Hilfe von Subsystem 1 alternative Stanzverfahren wie beispielsweise das Hohlschneiden, um die Schnittflächenqualität zu erhöhen. Hohlschneiden ist ein Verfahren, welches durch den Einsatz eines neuartigen Stempelkantendesigns erhöhte Glattschnittanteile erzielen kann [18].
- **Synthese:** Auf Basis physikalischer Modelle werden mit Subsystem 3 neue Designparameter (z. B. für den Schneidstempel) berechnet.
- **Interaktion:** Das System schließt den Zyklus mit einer quellenbasierten Ausgabe an den Bediener ab.

4.1 Experimenteller Aufbau und Methodik

Die Validierung des gewählten Ansatzes erfolgte durch einen systematischen Baseline-Vergleich, bei dem der konzipierte Stanz-Agent einem Zero-Shot[1]-LLM (ohne Tool-Anbindung und externen Wissenszugriff) gegenübergestellt wurde. Als LLM wurde ebenfalls Claude 4.5 Sonnet verwendet, jedoch ohne Möglichkeit zur Verwendung von zusätzlichen Werkzeugen. Die Bewertung der Antworten erfolgte durch Experten des Instituts für Umformtechnik. Als primäre Evaluationsmetriken dienten:

- die Halluzinationsrate bestimmt durch das Verhältnis von faktisch wahren zu falschen Sätzen in der Antwort und
- die Korrektheit physikalischer Berechnungen und Aussagen.

4.2 Ergebnisse der Testfälle

Testfall 1: Domänenspezifische Wissensabfrage

Im ersten untersuchten Testfall wurde die Wissensextraktion hinsichtlich der Frage „Welches Wirkprinzip steht hinter dem neuartigen Scherschneidverfahren Hohlschneiden?“ geprüft. Während das Zero-Shot-LLM zu oberflächlichen Generalisierungen und teils falschen Aussagen neigte, demonstrierte der Agent die Vorteile der RAG-basierten Architektur in Bezug auf sehr treffende Antworten. Durch den Zugriff auf die Fachliteratur konnte das Wirkprinzip präzise benannt und durch Zitationen belegt werden. Die Testergebnisse sind in Tabelle 1 zusammengefasst.

Tabelle 1 Vergleich Zero-Shot LLM vs. Stanz-Agent für Testfall 1

Modell	Zero-Shot LLM	Stanz-Agent
Halluzinationsanteil	Ca. 62%	0%
Bewertung durch Experten	Die gestellte Frage wurde nicht vollständig beantwortet. Ein Großteil der Antwort ist gegenteilig zum tatsächlichen Prozessverhalten.	Der Agent gibt die richtige Antwort.

Testfall 2: Interpretation von Prozessdaten

Gegenstand dieses zweiten Testfalls war die Analyse bereitgestellter Kraft-Weg-Verläufe. Die Kraftwegverläufe wurden beim Scherschneiden von DP800 (Stempeldurchmesser: 10mm, Blechdicke: 1mm, Schneidspalt 15%) aufgezeichnet. Ziel war es, auf Basis dieser Zeitreihen den an der Schnittfläche entstehenden Glattschnittanteil zu bestimmen. Da das Zero-Shot LLM grundsätzlich über keine Möglichkeit verfügt direkt auf solche Daten zuzugreifen, wurde die gemessene Zeitreihe in das Antwortfeld kopiert. Das LLM analysierte die Zeitreihe und schätzte den Glattschnitt auf Basis des aufgrund des auftretenden Kraftabfall. Dabei konnte das Modell einen Kraftabfall im Kraftverlauf richtig erkennen und prognostizierte auf Basis dessen einen Glattschnittanteil von

[1] Ausführung einer Aufgabe durch das Sprachmodell ohne vorherige Bereitstellung von Lösungsbeispielen im Eingabetext.

60-65%. Der agentische Ansatz hingegen nutzt das durch Subsystem 2 bereitgestellte Software-Werkzeug und bestimmte den Wert korrekt mit 25%.

Testfall 3: Synthese und Berechnung von Designparametern des Schneidstempels

Im finalen Testfall sollten die Designparameter eines Hohlstempels für eine spezifische Materialcharge (Blechdicke: 1mm, Werkstoff: DP 600) derart berechnet werden, dass ein gewünschter Glattschnittanteil von 71% erzielt wird. Dem Zero-Shot LLM wurden die notwendigen Prozessparameter mit einer Aufgabenbeschreibung übergegeben, jedoch war das Modell nicht in der Lage, die Optimierung durchzuführen und lieferte stattdessen halluzinierte Designparameter als Ergebnis. Diese halluzinierten Designparameter sind für die Auslegung eines Hohlstempels nicht anwendbar bzw. nicht technisch umsetzbar. Das agentische LLM greift auf das durch Subsystem 3 bereitgestellte Berechnungsskript zu und erfragte für die anschließende Berechnung spezifische Prozessparameter vom Anwender. Nach Eingabe dieser Parameter startete das Modell die Berechnung und gab die erforderlichen Designparameter aus. Zur Kontrolle wurden die bestimmten Parameter mittels einer kalibrierten Scherschneidsimulation überprüft. Der numerisch bestimmte Glattschnittanteil betrug 73,75%. Die Ergebnisse des Testfalls 3 sind in Tabelle 2 zusammengefasst.

Tabelle 2:Vergleich Zero-Shot LLM vs Stanz-Agent für Testfall 3

Modell	Halluzinationen	Abweichung zu Ground Truth	Bewertung durch Experten
Zero-Shot LLM	Halluzination von Prozessparametern	-	Das LLM halluziniert Prozessparameter und ist daher zur Auslegung nicht geeignet
Stanz-Agent	Keine Halluzinationen	Abweichung von ca. 3,7%	Auf Basis der bestimmten Parameter ist eine Werkzeugauslegung und anschließende Produktion möglich.

5. Diskussion des Konzeptes und der Testergebnisse

Im Folgenden werden das Konzept sowie die erzielten Testergebnisse einer kritischen Diskussion unterzogen. Ziel ist es, die Potenziale und Limitationen agentenbasierter Large Language Models (LLMs) zur Unterstützung von Operatoren im industriellen Umfeld aufzuzeigen.

5.1 Technische Validität und Funktionsweise

Die Vorteile des vorgestellten Assistenzsystems entstehen aufgrund der Integration externer Software-Tools. Im Gegensatz zu herkömmlichen LLMs, die zu Halluzination neigen, wird die Zuverlässigkeit des Systems durch die Auslagerung komplexer Berechnungen an spezialisierte Algorithmen sichergestellt. Dies gewährleistet eine hohe Reproduzierbarkeit der Schlussfolgerungen, da das agentische LLM lediglich die Orchestrierung übernimmt, während die eigentliche Datenverarbeitung innerhalb definierter Parameter erfolgt. Dies führt zu einer Reduktion der Fehlerrate im Vergleich zu rein stochastischen Modellen.

Der hier vorgestellte agentische Ansatz wurde primär für die Optimierungsebene eines Scherschneidprozesses konzipiert, die eine Prozessanpassung und -planung im Sekundenbereich oder

Minutenbereich realisiert. Die technologische Limitierung ergibt sich aus der Tatsache, dass die Abfrage der verwendeten High-Reasoning-Modelle über cloudbasierte Schnittstellen erfolgt. Dieser Abfrageprozess bedingt Latenzzeiten, die durch die Netzwerkkommunikation und die Rechenzeit der Modellanbieter entstehen. Daher ist dieser Ansatz für sicherheitskritische Steuerungsaufgaben, die Reaktionen im Millisekundenbereich erfordern, nicht geeignet. Solche Aufgaben müssen durch klassische Automatisierungstechnik wie speicherprogrammierbare Steuerungen (SPS), welche die unmittelbare Prozessstabilität gewährleisten, ausgeführt werden.

Potential zur Überwindung dieser Latenzbarriere bietet die zukünftige Integration lokaler Small Language Modelle (SLMs). Durch die Reduzierung der Parameteranzahl bei gleichzeitiger Spezialisierung auf domänenspezifische Aufgaben ermöglichen diese Modelle einen direkten On-Premise- oder sogar Edge-Betrieb unmittelbar an der Fertigungslinie. Die lokale Ausführung kann den Zeitaufwand für die Cloud-Kommunikation eliminieren und gleichzeitig die Datensouveränität sowie die Autonomie des Systems erhöhen.

5.2 Sicherheit, Datenschutz und Systemarchitektur

Ein zentraler Aspekt der hier vorgestellten Systemarchitektur für ein solches Assistenzsystem ist die Nutzung des MCPs als Abstraktionsschicht. Dies ermöglicht es, Merkmale und prozessrelevante Metadaten an das Modell zu übermitteln, während trotzdem sensible Produktionsdaten und das firmenspezifische Know-How innerhalb der lokalen Infrastruktur verbleiben. Die Modularität des Systems wird durch ein LLM-agnostisches Design gewährleistet. Durch die Plug-and-Play-Fähigkeit von MCP können neue Sensoreinheiten, Datenbanken oder zusätzliche Agenten für komplexe Fertigungslinien ohne grundlegende Überarbeitung der Architektur integriert werden. Diese Skalierbarkeit ist ein Vorteil für den Einsatz in hochdynamischen Produktionsumgebungen dar.

5.3 Soziotechnische Faktoren

Die Akzeptanz und Anwendung komplexer Systeme durch Operatoren ist unmittelbar an die Anwenderakzeptanz gekoppelt. Ein Faktor hierbei die Überwindung der Intransparenz durch erklärbare Ausgaben des Assistenzsystem. Diese Erklärbarkeit bildet die Grundlage für den Aufbau von Vertrauen in den Assistenten. Der Operator agiert nicht mehr nach Vorgaben einer „Black Box“, sondern kann die datenbasierten Entscheidungen verifizieren. Ermöglicht wird dies durch ein RAG-System, welches das präzise Zitieren der zugrunde liegenden Primärquellen ermöglicht. Indem das System Aussagen mit Verweisen auf Maschinendokumentationen, Normen oder historischen Prozessdaten verknüpft, wird hiermit eine höhere Ebene der Nachvollziehbarkeit erreicht. Die Aussage des Systems ändert sich dadurch von einer rein zufälligen Generierung hin zu einer evidenzbasierten Argumentation.

5.4 Wirtschaftliche und organisatorische Implikationen

Aus betriebswirtschaftlicher Perspektive kann der Einsatz agentischer LLMs zu einer Steigerung der Overall Equipment Effectiveness (OEE) führen. Diese Effizienzsteigerung wird primär durch

die Reduktion von Ausschuss sowie die Minimierung ungeplanter Stillstandzeiten realisiert, da das System bei Störungen unmittelbar evidenzbasierte Lösungsvorschläge liefert.

Vor dem Hintergrund des zunehmenden Fachkräftemangels stellt die Wissensdemokratisierung einen strategischen Wettbewerbsvorteil dar. Das in diesem Beitrag vorgestellte System ermöglicht eine skalierbare, ortsunabhängige sowie dauerhafte Bereitstellung von Expertenwissen direkt am Point-of-Use, wodurch Best Practices und Problemlösungsstrategien unabhängig von Schichtzyklen oder der Präsenz von Spezialisten verfügbar werden. Die Reduktion der Abhängigkeit von hochspezialisierten Wissensträgern mindert das Risiko von Wissensverlusten durch Mitarbeiterwechsel oder Renteneintritt. Gleichzeitig wird die Qualifizierung neuer Mitarbeiter beschleunigt, da das Assistenzsystem komplexe Prozesszusammenhänge situativ vermittelt und somit eine „On-the-Job"-Weiterbildung ermöglicht.

5.5 Limitationen des Operatorassistenzsystems

Trotz des erheblichen Potenzials zeigt die vorgestellte Architektur für ein Assistenzsystem auch einige spezifische Nachteile. Ein grundlegendes Problem besteht in der Gebundenheit an das Tool-Ökosystem. Der agentische Ansatz kann Problemlösungen nur innerhalb des Bereiches durchführen, für die bereits entsprechende Werkzeuge, Modelle oder Schnittstellen vorhanden sind. Die kognitive Kapazität des LLM dient lediglich der Orchestrierung, falls die technologische Infrastruktur zur physischen oder rechnerischen Umsetzung einer Aufgabe fehlt, sinkt die Wirksamkeit des Systems.

Zusätzlich ergeben sich aus der Systemstruktur weitere kritische Aspekte. Die Verlässlichkeit der Ergebnisse ist unmittelbar an die Güte der verwendeten Software-Werkzeuge und Datenbanken gekoppelt. Fehlerhafte Algorithmen oder veraltetes Wissen führen zwangsläufig zu schlechten oder sogar falschen Entscheidungen des Agenten. Haftungsrechtliche Fragestellungen bei Fehlentscheidungen von autonomen Prozesssteuerungen sind derzeit noch Gegenstand des Diskurses und erfordern klare Verantwortlichkeitsstrukturen im Betrieb.

6. Zusammenfassung und Ausblick

Angesichts steigender Qualitätsanforderungen und des drohenden „Brain-Drains" durch den demografischen Wandel präsentiert dieser Beitrag eine agentische Architektur, die Maschinenbediener bei der Überwachung von Scherschneidprozessen unterstützen soll. Durch die Verknüpfung eines agentischem LLM als zentralem Orchestrator mit speziellen Subsystemen wurde ein System geschaffen, das über MCP heterogene Datenquellen wie Wissensdatenbanken, Prozessanalytik und formalisiertes Expertenwissen verknüpfen kann.

Die experimentelle Validierung des hier vorgestellten Konzeptes an drei ausgesuchten Testfällen belegt, dass die gezielte Werkzeuganbindung die inhärenten Probleme klassischer LLMs überwinden kann. Im Gegensatz zu Zero-Shot-Ansätzen bietet das System eine robuste, quellenbasierte Entscheidungshilfe, die zur Prozessstabilisierung und Wissensdemokratisierung beiträgt.

Zukünftige Forschungsarbeiten sollten die Integration multimodaler Modelle adressieren, um durch die zusätzliche Verarbeitung von Bilddaten ggfs. visuelle Qualitätsmerkmale zu erschließen. Zur Steigerung der Datensouveränität und Reduktion der Latenz bietet zudem die Umstellung

auf Small Language Models (SLMs) Potenzial zur weiteren Verbesserung des Systems, in dem dadurch On-Premise- oder sogar Edge-Betrieb möglich sind. Darüber hinaus soll die Generalisierbarkeit des vorgestellten Ansatzes geprüft werden. Dabei steht die Übertragung der agentischen Architektur auf weitere Umform- und Fertigungsprozesse im Fokus, um das Potenzial einer domänenübergreifenden Assistenzplattform für die industrielle Produktion zu validieren.

Literatur

[1] [1]K. Lange, Ed. *Umformtechnik Handbuch für Industrie und Wissenschaft: Band 3: Blechbearbeitung*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1990.

[2] M. Polanyi, *Personal knowledge: Towards a post-critical philosophy*. London: Routledge & Kegan Paul, 1958.

[3] Ulmer *et al., Der demografische Wandel und seine Folgen für die Sicherstellung des Fachkräftenachwuchses* (Wissenschaftliche Diskussionspapiere 106). Bonn: BIBB, 2008. [Online]. Available: http://nbn-resolving.de/urn:nbn:de:0035-0290-0

[4] Y. Liu, T. Ji, X. Guo, X. Xu, and J. Polzer, "A survey of cognitive digital twin and the potential use of LLMs," *Manufacturing Letters*, vol. 44, pp. 1242–1253, 2025, doi: 10.1016/j.mfglet.2025.06.144.

[5] M. Besta *et al.,* "Reasoning Language Models: A Blueprint," doi: 10.48550/arXiv.2501.11223.

[6] A. Vaswani *et al.,* "Attention Is All You Need," doi: 10.48550/arXiv.1706.03762.

[7] Open AI. "ChatGPT ist da." [Online]. Available: https://openai.com/de-DE/index/chatgpt/

[8] Z. Ji *et al.,* "Survey of Hallucination in Natural Language Generation," *ACM Comput. Surv.*, vol. 55, no. 12, pp. 1–38, 2023, doi: 10.1145/3571730.

[9] H. Zhao *et al.,* "Explainability for Large Language Models: A Survey," 2023, doi: 10.48550/arXiv.2309.01029.

[10] A. Plaat, M. van Duijn, N. van Stein, M. Preuss, P. van der Putten, and K. J. Batenburg, "Agentic Large Language Models, a survey," 2025, doi: 10.48550/arXiv.2503.23037.

[11] Anthropic AI. "Introducing the Model Context Protocol." Accessed: Jan. 16, 2025. [Online]. Available: https://www.anthropic.com/news/model-context-protocol

[12] X. Hou, Y. Zhao, S. Wang, and H. Wang, "Model Context Protocol (MCP): Landscape, Security Threats, and Future Research Directions," 2025, doi: 10.48550/arXiv.2503.23278.

[13] M. Becker, P. Niemietz, and T. Bergs, "Explainable neural network for time series-based condition monitoring in sheet metal shearing," *J Intell Manuf*, 2025, doi: 10.1007/s10845-025-02596-3.

[14] C. Kubik, D. A. Molitor, M. Rojahn, and P. Groche, "Towards a real-time tool state detection in sheet metal forming processes validated by wear classification during blanking," *IOP Conf. Ser.: Mater. Sci. Eng.*, vol. 1238, no. 1, p. 12067, 2022. doi: 10.1088/1757-899X/1238/1/012067. [Online]. Available: https://iopscience.iop.org/article/10.1088/1757-899X/1238/1/012067

[15] P. Lewis *et al.,* "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," doi: 10.48550/arXiv.2005.11401.

[16] A. Schenek, M. Görz, M. Liewald, and K. R. Riedmüller, "Prediction of Cutting Surface Parameters in Punching Processes Aided by Machine Learning," in *TMS 2023 152nd Annual*

Meeting & Exhibition Supplemental Proceedings (The Minerals, Metals & Materials Series), Cham: Springer Nature Switzerland, 2023, pp. 607–619.

[17] M. Görz, M. Liewald, K. R. Riedmueller, and A. Schenek, "Optimierung des Wissenstransfers durch KI und Explainable AI," in *Tagungsband T 56: Innovationsstandort Europa: Effiziente Blechverarbeitung: Pressen, Systeme, Prozesse*, Würzburg, Europäische Forschungsgesellschaft für Blechverarbeitung, Ed., 2025.

[18] A. Schenek, S. Senn, and M. Liewald, Erhöhung der Schnittflächenqualität mittels Hohlschneiden (EFB-Forschungsbericht Nr. 593). Hannover: Europäische Forschungsgesellschaft für Blechverarbeitung e.V, 2023.

Verlag Meisenbach GmbH, Geisfelder Straße 14, 96050 Bamberg/Deutschland
http://www.umformtechnik.net/whitepaper/