GfA, Sankt Augustin (Hrsg.): Frühjahrskongress 2026, Kassel
Menschengerechte Arbeitsgestaltung

KI-gestütztes Open-Source-Tool zur Effizienzsteigerung im Lern- und Wissensmanagement

Jannis PFISTER[1], Matthias JENNER[1], Till GÜNTHER, Rainer MÜLLER[2]

[1]ZeMA – Zentrum für Mechatronik und Automatisierungstechnik gGmbH, Eschberger Weg 46, D-66121 Saarbrücken
[2] Lehrstuhl für Montagesysteme, Universität des Saarlandes, Eschberger Weg 46, D-66121 Saarbrücken

Kurzfassung: Angesichts von Fachkräftemangel und demografischem Wandel wird die Sicherung von Erfahrungs- und Prozesswissen für KMU zum kritischen Wettbewerbsfaktor. Die Überführung dieses Wissens in digitale Lernprozesse scheitert jedoch oft an begrenzten Ressourcen. KI-gestützte Werkzeuge bieten hierfür erhebliche Effizienzpotenziale, erfordern jedoch strikte Datensouveränität und Transparenz. Der vorgestellte Ansatz beschreibt eine Open-Source-Architektur, die lokal betriebene Large Language Models (LLM) mit einer dynamischen Anbindung an interne Wissensquellen kombiniert, um SCORM-kompatible Lerninhalte effizient zu generieren. Ein integrierter Human-in-the-Loop-Prozess gewährleistet dabei die fachliche Validierung der Inhalte durch Experten. Die Evaluation des Prototyps belegt signifikante Effizienzsteigerungen und eine deutliche Beschleunigung des Erstellungsprozesses, verdeutlicht jedoch zugleich, dass die didaktische Qualität ohne menschliche Nachbearbeitung schwanken kann. Der Beitrag zeigt damit einen praxisnahen, aber differenzierten Weg zur Integration generativer KI in bestehende Lern- und Wissensprozesse auf.

Schlüsselwörter: Wissensmanagement, Künstliche Intelligenz, Human-in-the-Loop, Lernmanagementsysteme, Open Source, KMU

1. Einleitung & Motivation

Wissen gilt in KMU als zentrale strategische Ressource, wird in der Praxis jedoch häufig unsystematisch und personenabhängig gehandhabt (Durst et al. 2022; Escobar-Castillo & Velandia-Pacheco 2024). Vor allem Erfahrungs- und Prozesswissen bleibt an einzelne Personen gebunden und ist nur begrenzt dokumentiert. Damit stellt sich nicht nur die Frage, wie dieses Wissen erhalten werden kann, sondern auch, wie es so aufbereitet wird, dass es in kontinuierlichen Lern- und Qualifizierungsprozessen wirksam genutzt werden kann. Lernmanagementsysteme bieten hierfür eine technische Grundlage, ihre Wirkung hängt jedoch auch unmittelbar von der Verfügbarkeit passender und aktueller Inhalte ab, deren Erstellung in KMU als wesentlicher Engpass identifiziert werden kann (Oni-Orisan 2016).

Generative KI und große Sprachmodelle eröffnen die Möglichkeit, Wissensmanagement und Lernen enger zu verzahnen, indem vorhandenes Fachwissen effizient in strukturierte Lern- und Wissensbausteine überführt wird (Li et al. 2025). Erste Arbeiten zeigen Effizienzpotenziale, verweisen aber zugleich auf Risiken wie Halluzinationen, begrenzte fachliche Validität und fehlende Transparenz der Modelle (Cossio 2024). Für KMU mit sensiblen Daten rücken daher

datensouveräne, lokal betreibbare Lösungen und explizite Qualitätsprozesse in den Vordergrund. Human-in-the-Loop-Ansätze, in denen automatisch erzeugte Inhalte systematisch durch Fachexpertinnen und Fachexperten geprüft und freigegeben werden, werden in diesem Zusammenhang als zentraler Baustein einer verantwortungsvollen Nutzung betrachtet (Barberá 2023).

Vor diesem Hintergrund wurde ein Prototyp einer KI-gestützten Open-Source-Architektur realisiert, der lokale LLMs, interne Wissensquellen und bestehende Lernmanagementsysteme verbindet, um die Erstellung von Lern- und Wissensinhalten sowohl effizienter als auch kontrollierbar sowie steuerbar zu gestalten. Im Mittelpunkt stehen die Fragen, wie KI zuverlässig in die Content-Erstellung eingebunden werden kann, welche Rolle der Mensch im Rahmen von Human-in-the-Loop-Prozessen für Qualität und Akzeptanz spielt und inwieweit lokal betriebene LLM-Architekturen eine praxistaugliche Option für KMU darstellen.

2. Technologische Grundlagen und Stand der Technik

Für das Verständnis der nachfolgend beschriebenen Systemarchitektur ist eine präzise Einordnung der verwendeten Technologien und Standards unerlässlich. Die Basis der semantischen Verarbeitung bilden Large Language Models (LLM), die hierbei nicht als statische Datenbank, sondern als „kognitiver Motor“ fungieren, um unstrukturierte Informationen zu analysieren und in neue Formate zu transferieren (Kalweit & Gabriel 2024). Ein wesentliches Defizit generativer Modelle ist jedoch ihre Neigung zu Halluzinationen, weshalb für den industriellen Einsatz der Ansatz der Retrieval-Augmented Generation (RAG) zwingend erforderlich ist. Hierbei generiert das Modell Antworten nicht frei, sondern ausschließlich auf Basis injizierter Firmendokumente (Lewis et al. 2020). Damit das System diese Dokumente kontextbezogen identifizieren kann, kommt die Vektorisierung (Embeddings) zum Einsatz. Ein Embedding-Modell transformiert Texte in mathematische Vektoren, sodass Inhalte nach ihrer semantischen Bedeutung und nicht nur nach Schlagworten gefunden werden (Reimers & Gurevych 2019).

Eine zentrale Herausforderung für KMU stellt hierbei die Datensouveränität dar, da die Nutzung cloudbasierter Modelle Risiken wie Modell-Inversion oder Datenabfluss birgt (Barberá 2025). Ergänzend hierzu führt die Anbindung externer Cloud-Dienste zu einer kritischen operativen Abhängigkeit von der Verfügbarkeit fremder Infrastrukturen. Netzwerkausfälle oder Änderungen an API-Schnittstellen seitens der Anbieter können die Betriebsfähigkeit des Systems unmittelbar gefährden. Um diese Sicherheits- und Verfügbarkeitslücke zu schließen, ist eine Lokale Inferenz (On-Premise) notwendig, bei der die gesamte Datenverarbeitung auf eigener Hardware innerhalb des geschützten Unternehmensnetzwerks stattfindet.

Da etablierte Lernmanagementsysteme (LMS) wie Moodle (Dougiamas & Taylor 2003) primär als administrative Plattformen dienen und nach aktuellem Stand der Technik keine native Möglichkeit besitzen, komplexe Kursstrukturen vollautomatisiert und RAG-basiert aus lokalen Wissensquellen zu generieren, entsteht eine funktionale Lücke im digitalen Workflow. Zwar existieren Standards wie SCORM für den Austausch von Lerninhalten (ADL 2009), doch fehlt es an integrierten Werkzeugen, die diesen Standard unter Berücksichtigung von Datensouveränität und KI-gestützter Inhaltserstellung bedienen. Sowohl Sicherheitslücken in Cloud-Lösungen als auch die fehlende Faktentreue ohne RAG und unzureichende LMS-Tools begründen den Bedarf für diese Eigenentwicklung. Um eine praxistaugliche Anwendung zu schaffen, ist eine dedizierte Architektur erforderlich, die diese Aspekte modular verknüpft.

3. Systemarchitektur und Implementierung

Die Realisierung der Anforderung nach Datensouveränität bei gleichzeitiger Nutzung moderner Sprachmodelle erfolgt durch die in Abbildung 1 dargestellte modulare, containerisierte Architektur.

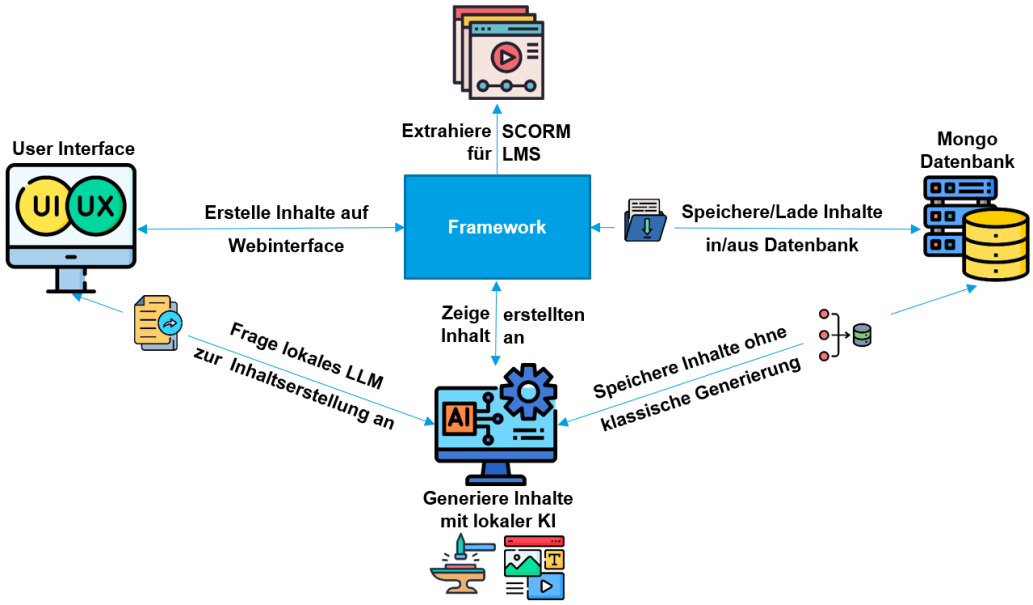

***Abbildung 1**: Arbeitsweise und Kommunikation des KI-gestützten Open-Source-Tools mit beteiligten Komponenten*

Das System ist so konzipiert, dass es vollständig "On-Premise" auf Standard-Serverhardware in KMU betrieben werden kann. Das Herzstück bildet die Interaktion zwischen der Benutzeroberfläche und den Inferenz-Modellen. Für die Verwaltung der lokalen Sprachmodelle wird Ollama als Backend-Engine eingesetzt. Für die konkrete Implementierung wurde das Open-Source-Modell Mistral 7B gewählt. Das Modell zeichnet sich durch ein optimales Verhältnis zwischen sprachlicher Leistungsfähigkeit im Deutschen und moderatem Ressourcenbedarf aus. In der Folge wird der lokale Betrieb auf Edge-Geräten am Shopfloor wirtschaftlich ermöglicht (Jiang et al. 2023).

Die Kommunikation wird durch eine REST-API initiiert, die durch OpenWebUI als Frontend und Management-Layer gesteuert wird. Eine Besonderheit der Implementierung ist das differenzierte Rechtemanagement über API-Schlüssel. Da verschiedene Nutzergruppen Zugriff auf unterschiedliche Wissensbestände benötigen, erlaubt die Architektur die Vergabe individueller Schlüssel, die steuern, auf welche Vektor-Indizes das RAG-System zugreifen darf. Für die Speicherung des Zustands von generierten Kursinhalten wird die dokumentenorientierte Datenbank MongoDB eingesetzt. Im Gegensatz zu relationalen Datenbanken erlaubt ihr JSON-basiertes Schema die flexible Speicherung der heterogenen Datenstrukturen, die bei der Generierung von Lerninhalten entstehen. Die generierten Rohdaten werden an dieser Stelle direkt persistiert, um eine Versionierung und spätere Bearbeitung zu ermöglichen.

Da Lernmanagementsysteme wie Moodle zwar erste Ansätze zur KI-Textgenerierung bieten, jedoch keine native Möglichkeit besitzen, komplexe Kursstrukturen RAG-basiert sowie vollautomatisiert aus lokalen Wissensquellen zu

generieren, fungiert das Adapt Learning Framework als notwendige Middleware. Der Prozess läuft teilautomatisiert ab: Über die API fordert das Adapt-Tool Inhalte beim lokalen LLM an. Zunächst werden diese im Editor verarbeitet, wo der Ansatz des "Human-in-the-Loop" zum Tragen kommt. In diesem Prozess überprüft und korrigiert der Autor die KI-Vorschläge. Nach erfolgter Freigabe kompiliert das Framework die Inhalte in ein standardisiertes SCORM-Paket. Der finale Import in das LMS (Moodle) erfolgt bewusst als Datei-Upload. Dieser Medienbruch fungiert als zusätzliches "Quality Gate": Der Kurs wird nicht ungeprüft beim Lerner platziert, sondern muss aktiv durch einen Administrator freigegeben werden. Moodle fungiert in der Folge als revisionssichere "Single Source of Truth" für den Lernfortschritt.

4. Operativer Workflow und Funktionalität

Die technische Architektur wird durch einen definierten Workflow operationalisiert, der die Stärken der generativen KI mit der notwendigen menschlichen Kontrollinstanz verbindet. Die Anbindung des lokalen LLMs über eine offene API-Schnittstelle bietet dabei einen entscheidenden Vorteil gegenüber einer monolithischen Integration, da sie perspektivisch auch Modalitäten wie Audio-Transkription einbinden kann. Der Fokus des aktuellen Prototyps liegt auf der wissensbasierten Texterstellung und der automatisierten Quiz-Generierung. Bei der Texterstellung besteht die Möglichkeit, Inhalte wahlweise angeleitet oder per "Free Prompt" auf Basis vektorisierter Firmendokumente zu generieren. Das Quiz-Modul hingegen erlaubt die automatische Ableitung von Testfragen inklusive Bewertungsschema direkt in die Datenbankstruktur.

Für die Integration dieser Inhalte in das Lernmanagementsystem wurde bewusst auf eine unmittelbare, vollautomatisierte Injektion in die Moodle-Kursdatenbank verzichtet. Die interne Struktur von Moodle-Kursen ist komplex und versionsabhängig, sodass ein direkter Schreibzugriff die Datenintegrität gefährden kann. Stattdessen fungiert das Adapt Framework als eine stabilisierende Zwischenschicht, in der Inhalte strukturiert und finalisiert werden, bevor sie als gekapseltes SCORM-Paket exportiert werden. Dieser Workflow bedingt systemseitig eine menschliche Validierung (Human-in-the-Loop). Obwohl KI-Modelle durch die RAG-Anbindung zu Wiederholungen oder fachlichen Unschärfen neigen können, müssen die im Editor generierten Vorschläge durch Fachexperten aktiv geprüft, gekürzt und freigegeben werden. Der Mensch übernimmt somit die finalisierende Bearbeitung des von der KI gelieferten Rohbaus, was die Verbreitung fehlerhaften Wissens am Shopfloor verhindert und die Akzeptanz der Lösung erhöht.

5. Evaluation und Ergebnisse

Zur Validierung der entwickelten Systemarchitektur wurde ein zweistufiges Evaluationskonzept umgesetzt, das sowohl die Effizienz des Erstellungsprozesses als auch die Lernerfahrung der Nutzer untersuchte. Als Testszenarien dienten die realen Industriethemen „Ergonomie am Arbeitsplatz“ und „Kreislaufwirtschaft“.

Im ersten Schritt erfolgte ein quantitativer A/B-Vergleich der Erstellungsdauer zwischen der konventionellen manuellen Ausarbeitung und der KI-gestützten Erstellung mittels der lokalen RAG-Architektur. Dieser Vergleich wurde exemplarisch von zwei Fachexperten durchgeführt, um das Einsparpotenzial bei der Content-Erstellung zu ermitteln. Die Messungen belegen eine deutliche Beschleunigung durch

das Tool: Während die manuelle Erstellung des Ergonomie-Kurses 20 Stunden beanspruchte, konnte dieser Wert durch die KI-Pipeline auf 8 Stunden reduziert werden, was einer Effizienzsteigerung von 60 Prozent entspricht. Ähnliche Ergebnisse zeigten sich beim Thema Kreislaufwirtschaft mit einer Reduktion von 15 auf 7 Stunden. Die hierdurch freigewordenen Ressourcen ermöglichen eine Verlagerung der Arbeitszeit von der reinen Content-Erzeugung hin zur qualitativen Veredelung.

Auf Basis dieser generierten Inhalte wurde im zweiten Schritt die Qualität und Akzeptanz evaluiert. Hierfür untersuchte eine Testgruppe von acht Probanden (N=8) aus dem akademischen und industriellen Umfeld (Durchschnittsalter 30 Jahre) die Kurse in einer Blindstudie. Die Auswertung der Feedback-Daten zeigt eine differenzierte Wahrnehmung der KI-Inhalte. Grundsätzlich signalisierte die Mehrheit der Teilnehmenden eine hohe Offenheit gegenüber der Technologie; die Aussage, dass KI genutzt werden darf, solange die Qualität stimmt, erhielt hohe Zustimmungswerte. Jedoch korrelierte die Lernmotivation stark mit dem Grad der menschlichen Nachbearbeitung. In den Freitext-Rückmeldungen wurde mehrfach auf „Redundanzen“ und eine als „langatmig“ empfundene Textstruktur hingewiesen, sofern Inhalte ungeprüft übernommen wurden. Insbesondere bei generischen Prompts tendierte das Modell zu weitschweifigen Formulierungen, die von den Probanden als „ermüdend“ charakterisiert wurden. Zudem fiel auf, dass bei unzureichender Validierung Distraktoren in Multiple-Choice-Tests teilweise unlogisch generiert wurden, was die didaktische Glaubwürdigkeit minderte.

Trotz dieser qualitativen Mängel in der Rohfassung bestand unter den Testpersonen keine prinzipielle Ablehnung, was sich auch nochmals in Abbildung 2 deutlich zeigt. Dargestellt ist die Zustimmung auf einer 7-stufigen Likert-Skala (1 = „Stimme überhaupt nicht zu“ bis 7 = „Stimme voll und ganz zu“)

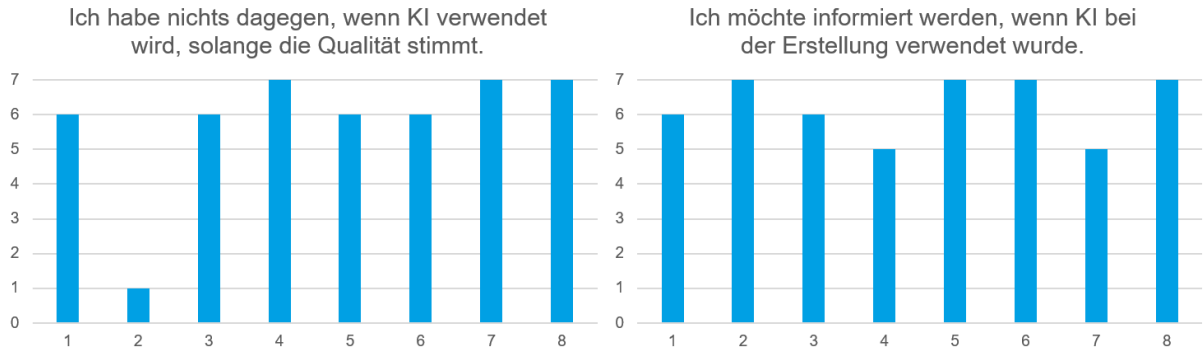

Abbildung 2: *Vergleich der Nutzerakzeptanz zur Forderung nach Transparenz (N=8)*

Die quantitative Auswertung bestätigt vielmehr, dass der Wunsch nach Transparenz und menschlicher Kontrolle dominiert: Die Probanden forderten nahezu einstimmig, dass KI-generierte Inhalte transparent gekennzeichnet und zwingend durch Fachexperten kuratiert werden müssen. Die Studie belegt somit empirisch, dass die Akzeptanz von KI-Lernmedien weniger von der Technologie selbst, als vielmehr von der Sichtbarkeit der Qualitätssicherung abhängt

6. Fazit und Ausblick

Der realisierte Prototyp demonstriert, dass der Einsatz lokaler Large Language Models eine leistungsfähige und datensouveräne Lösung für die Herausforderungen des Wissensmanagements in KMU darstellt. Die Evaluation zeigt eine signifikante

Steigerung der Effizienz bei der Erstellung von SCORM-basierten Lerninhalten um bis zu 60 Prozent, wobei keine sensible Prozessdaten die unternehmenseigene Infrastruktur verlassen. Gleichzeitig verdeutlichen die Ergebnisse, dass Technologie allein keine Garantie für didaktische Qualität ist. Der etablierte Human-in-the-Loop-Prozess hat sich als unverzichtbare Instanz zur fachlichen Validierung erwiesen, um Redundanzen zu vermeiden und die Akzeptanz bei den Lernenden zu sichern. In diesem System nimmt der Mitarbeiter nicht mehr die Rolle eines einfachen Autors ein, sondern fungiert als qualitätssichernder Gestalter. Der Fokus zukünftiger Erweiterungen liegt primär auf dem Abbau von Zugangsbarrieren bei der Wissenserfassung. In Bezug auf die genannte Thematik wurde die Integration von Speech-to-Text-Schnittstellen in die Konzeptualisierung einbezogen. Das Ziel dieser Integration besteht darin, den Mitarbeitenden die Möglichkeit zu bieten, ihr Erfahrungswissen in einer niederschwelligen mündlichen Form einzubringen. Das System wird dahingehend erweitert, dass es neben der Verarbeitung von Texten auch für die Interpretation und Einbindung von technischen Zeichnungen, Bildmaterial sowie Videosequenzen in den Lernkontext vorgesehen ist.

7. Literatur

Adapt Learning Project (2024) Adapt Authoring Tool & Framework Documentation. https://www.adaptlearning.org. Zugegriffen: 15. Dez. 2025.

ADL, Advanced Distributed Learning Initiative (2009) SCORM 2004 4th Edition – Content Aggregation Model (CAM). Alexandria, VA: ADL Technical Team.

Barberá I (2025) AI Privacy Risks & Mitigations - Large Language Models (LLMs). European Data Protection Board (Support Pool of Experts Programme).

Cossio M (2025) A comprehensive taxonomy of hallucinations in Large Language Models. arXiv preprint arXiv:2508.01781.

Dougiamas M & Taylor PC (2003) Moodle: Using Learning Communities to Create an Open Source Course Management System. In: Lassner D & McNaught C (Eds) Proceedings of ED-MEDIA 2003. Honolulu: AACE, 171-178.

Durst S, Foli S & Edvardsson IR (2022) A systematic literature review on knowledge management in SMEs: current trends and future directions. Management Review Quarterly 73:1-35.

Escobar-Castillo A & Velandia-Pacheco G (2024) Categorizing the effects of knowledge management practices on SMEs: a literature review. TEC Empresarial 18:23-42.

Jiang AQ, Sablayrolles A, Mensch A et al. (2023) Mistral 7B. arXiv preprint arXiv:2310.06825.

Kalweit M & Gabriel (2024) Warum wir neu lernen müssen, mit Maschinen zu sprechen – eine Momentaufnahme der generativen KI. Ordnung der Wissenschaft 2:125-138.

Lewis P, Perez E, Piktus A et al. (2020) Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In: Larochelle H et al. (Eds) Advances in Neural Information Processing Systems (NeurIPS) 33. Red Hook, NY: Curran Associates, 9459-9474.

Li H, Fang Y, Zhang S et al. (2025) ARCHED: A Human-Centered Framework for Transparent, Responsible, and Collaborative AI-Assisted Instructional Design. Proceedings of Machine Learning Research 273:1-11.

Oni-Orisan A (2016) Knowledge management barriers in SMEs. In: Proceedings of the 3rd International Conference on Knowledge Management.

Reimers N & Gurevych I (2019) Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: Inui K et al. (Eds) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Hong Kong: Association for Computational Linguistics, 3982-3992.

Danksagung: Diese Forschungsarbeit entstand im Projekt KOMATRA (Förderkennzeichen 02L22C005), gefördert vom Bundesministerium für Bildung und Forschung (BMBF).