JG. 8 (2026), H. 1
DOI: 10.18716/OJS/MIDU/2026.1.10
VERÖFFENTLICHT AM 27.07.2026

M i D U
Medien im Deutschunterricht

HERMETISCHE KI – HERMENEUTISCHE HERAUSFORDERUNGEN

Matthias Leichtfried
Universität Wien | matthias.leichtfried@univie.ac.at

ABSTRACT

Wie kann man generative KI verstehen? Ausgehend von dieser Leitfrage problematisiert der Beitrag das Verstehen von KI im Kontext der Deutschdidaktik und der Auseinandersetzung mit generativer KI im Unterricht. Dazu werden drei zentrale Herausforderungen herausgearbeitet: Erstens Fehlvorstellungen über die Funktionsweise und Verlässlichkeit generativer KI; zweitens die Frage nach Wissensquellen, Selektionspraktiken und Autorität in einem dynamischen, teils widersprüchlichen Forschungsdiskurs; drittens die Blackbox-Problematik aktueller Systeme, die Nachvollziehbarkeit und kausale Erklärungen begrenzt. Anschließend diskutiert der Beitrag drei Ansätze zum Umgang damit (Interdisziplinarität; kulturwissenschaftliche Reflexion; epistemische Offenheit) und argumentiert, dass ein erweitertes, kulturwissenschaftlich reflektiertes Verstehen dieser Technologie gerade Aufgabe der Deutschdidaktik als eingreifender Kulturwissenschaft ist.

SCHLAGWÖRTER

KÜNSTLICHE INTELLIGENZ — LARGE LANGUAGE MODELS — DEUTSCHDIDAKTIK — KULTURWISSENSCHAFT — HERMENEUTIK — AI LITERACY

JG. 8 (2026), H. 1
DOI: 10.18716/OJS/MIDU/2026.1.10
VERÖFFENTLICHT AM 27.07.2026

M i D U

Medien im Deutschunterricht

ABSTRACT (ENGLISH)

HERMETICAL AI – HERMENEUTICAL CHALLENGES

How can generative AI be understood? Based on this key question, the article examines the understanding of AI in the context of German language teaching and generative AI in the classroom. Three central challenges are identified: First, misconceptions about how generative AI works and its reliability; second, the question of knowledge sources, selection practices, and authority in a dynamic, sometimes contradictory research discourse; third, the black box problem of current systems, which limits traceability and causal explanations. The article then discusses three approaches to dealing with this (interdisciplinarity; cultural studies reflection; epistemic openness) and argues, that an expanded, reflective understanding of this technology stemming from cultural studies is precisely the task of German didactics as an intervening cultural science.

KEYWORDS

— ARTIFICIAL INTELLIGENCE — LARGE LANGUAGE MODELS — GERMAN LANGUAGE EDUCATION — CULTURAL STUDIES — HERMENEUTICS — AI LITERACY

1 — EIN UNBEHAGEN – ERGEBNISSE EINER EXPLORATIVEN STUDIE

Wie verlässlich sind Chatbots, wenn es darum geht, Feedback zu geben? Wie sehr kann man also ChatGPT und anderen KI-Tools in Bezug auf Rückmeldung und Textkorrekturen vertrauen? Diese Fragen waren Ausgangspunkt einer Untersuchung (Engelmayr-Hofmann / Leichtfried 2026), bei der mittels eines konstruierten Textes voller orthografisch und grammatikalisch korrekter Grenzfälle (laut Variantenwörterbuch) in österreichischer Standardvarietät überprüft wurde, wie ChatGPT und andere Tools mit den verschiedenen Varietäten des Deutschen umgehen. Mit der Aufforderung, den Text auf sprachnormative Richtigkeit zu prüfen, gaben fast alle getesteten Tools an, dass Verstöße gegen ‚die Norm' des Deutschen vorlägen. Die Hypothese, dass Large Language Modelle ihre sprachlichen Normen nicht aufgrund linguistischer Expertise, sondern aufgrund der verwendeten (fehlerbehafteten) Daten generieren, konnte so untermauert werden. Auffällig war aber an den Ergebnissen nicht nur die problematische Normsetzung, sondern auch, dass in den jeweils 25 Durchläufen, mit denen jedes Modell getestet wurde, eine beträchtliche Varianz in den Antworten festzustellen war. Die Standardabweichung für das Modell GPT 4o-mini betrug beispielsweise 1,61. Eine mit KI nicht vertraute Person könnte jetzt die naive Frage stellen, ob ein Computer nicht immer zu dem gleichen Ergebnis kommen sollte und ob sich die Rechenmaschinen nicht gerade durch Konsistenz und Regelhaftigkeit von menschlichen Denkprozessen unterscheiden. Eine mit KI vertraute Person würde in diesem Fall sehr wahrscheinlich zu einer Erklärung ansetzen, in der die Stichwörter „Wahrscheinlichkeit" (Wolfram 2023), „Mustererkennung" (Nassehi 2019, 229) und möglicherweise sogar „temperature" (Peeperkorn et al. 2024) mit großer Sicherheit vorkommen. Tatsächlich funktionieren sogenannte künstliche neuronale Netzwerke nicht linear, sondern sie basieren auf stochastischen Prinzipien, weshalb der gleiche Prompt zu völlig unterschiedlichen Ergebnissen führen kann. Verwenden Nutzer:innen nun KI-Tools, um ihre Aufsätze überprüfen zu lassen, wäre es vermutlich hilfreich, zu verstehen, wie generative KI funktioniert, also beispielsweise über das Wissen um das stochastische Prinzip neuronaler Netzwerke zu verfügen.

So erklärt die Gesellschaft für Informatik in der Dagstuhlerklärung aus 2016, dass Schüler:innen befähigt werden sollen, selbstbestimmt mit digitalen Systemen umzugehen. Dies erfordere, „sie zu verstehen, zu erklären, im Hinblick auf Wechselwirkungen mit dem Individuum und der Gesellschaft zu bewerten sowie ihre Einflussmöglichkeiten zu sehen und nicht nur ihre Nutzungsmöglichkeiten zu kennen" (Gesellschaft für Informatik 2016, 2). In Bezug auf diese Ziele werden daher drei Perspektiven vorgeschlagen (anwendungsbezogene, gesellschaftlich-kulturelle, technologische), wobei das Verstehen und Erklären mit der technologischen Perspektive und der damit verbundenen Frage „Wie funktioniert das?" korrespondiert. Diese Perspektive schaffe „die technologischen Grundlagen und Hintergrundwissen für die Mitgestaltung der digitalen vernetzten Welt" (ebd.). Das Ziel des Verstehens und Erklärens gilt dann auch in Bezug auf generative KI als digitales System und es verwundert nicht, dass gängige Beschreibungen einer KI-Kompetenz, wie jene von Long und Magerko, das Wissen über die Technologie als zentralen Baustein dafür setzen (vgl. 2020, 4). Die technologische Perspektive gilt

nun aber nicht nur für den Informatik-Unterricht, sondern wird auch in jenen Fächern zum Ziel, in denen generative KI im Sinne der Ziel-Mittel-Ambivalenz sowohl als Gegenstand als auch anwendungsbezogen-methodisch als „Tool" eingesetzt wird. Dies gilt nicht zuletzt für den Deutschunterricht – wie ein wachsender deutschdidaktischer Forschungsdiskurs in Bezug auf generative KI und ihren Einsatz im Deutschunterricht nahelegt (Krammer / Leichtfried 2024; Fürstenberg / Müller 2024; Hesse / Jagdschian 2026 i.E.; Standke 2024; Steinhoff 2023; Birk 2023). Nun ist aber „Verstehen" von künstlicher Intelligenz im landläufigen Sinne keineswegs unproblematisch: So stellen Long und Magerko Wissenslücken im Bereich der künstlichen Intelligenz fest und sie kritisieren die daraus entstehenden „‚folk theories'" (Long / Magerko 2020, 4) über die Funktionsweise. Darüber hinaus ist man im Forschungsbereich Künstliche Intelligenz mit einer hohen Zahl an Publikationen konfrontiert, die auf einen unübersichtlichen und kontroversen Diskurs hindeuten.[1] Dies wirft im Bereich der Deutschdidaktik einige Fragen auf: Wie verstehen Didaktiker:innen, Schüler:innen und Lehrer:innen jene KI-Tools, mit denen sie arbeiten? Welches Wissen über die Funktionsweise ist ausreichend? Wie detailliert müssen Deutschdidakter:innen, Lehrpersonen und Schüler:innen um den Diskurs zu generativer KI und ihrer Funktionsweise Bescheid wissen? Welche Wissensquellen sind geeignet, um den aktuellen Stand der Forschung abbilden zu können? In der Fachdidaktik Deutsch sowie der Professionalisierungsforschung müssen daher neben Überzeugungen (beliefs), subjektiven Theorien und conceptions (vgl. Reusser / Pauli 2014) auch grundsätzliche Herausforderungen in Bezug auf das Verstehen von generativer KI und ihrer Funktionsweise in den Blick geraten. Der hier vorliegende Beitrag setzt sich zum Ziel, diese Herausforderungen zu problematisieren, um anschließend ein erweitertes Konzept von Verstehen im Zusammenhang mit deutschdidaktischer Forschung sowie mögliche Ansätze für den Umgang mit diesen mit den genannten Problemfeldern vorzuschlagen.

2 — KÜNSTLICHE INTELLIGENZ VERSTEHEN – DREI HERAUSFORDERUNGEN

Der vorliegende Beitrag wird die oben aufgeworfenen Fragen nicht beantworten können, sondern er soll anhand ihrer Problematisierung zeigen, dass die Auseinandersetzung mit der Funktionsweise generativer künstlicher Intelligenz sowohl in der Deutschdidaktik als auch im Deutschunterricht einerseits zwar notwendig ist, andererseits aber mit der Herausforderung verbunden ist, ein komplexes, technisches System und seine Funktionsweise verstehen zu müssen. Im Folgenden werden daher zunächst drei Herausforderungen in Bezug auf das Verstehen der Funktionsweise von generativer KI dargelegt.

2.1 HERAUSFORDERUNG FEHLVORSTELLUNGEN

Künstliche Intelligenz ist schon als Begriff an sich problematisch, weil nicht nur die dahinterliegenden Konzepte der Künstlichkeit und der Intelligenz keineswegs klar definierbar sind, sondern weil unter diesem Begriff zahlreiche unterschiedliche Kon-

1 Bereits 2023 zeigen Krenn et al. (2023), dass die Zahl der KI-bezogenen Paper exponentiell wächst, in einer Analyse der Publikationen zu KI im Zeitraum der ersten 43 Tage des Jahres 2026 zählt Thompson (2026) 6354 veröffentlichte KI-bezogene Paper auf arXiv.org.

zepte, Technologien und Anwendungen gefasst werden. Koehler (2024) verweist in ihrem Beitrag zur Begriffsbestimmung auf einen Bericht von Legg und Hutter aus dem Jahr 2007, bei dem über 70 Definitionen gesammelt wurden, und sie konzediert, dass die Fähigkeit, von einer Maschine selbstgestellte Ziele zu erreichen, als häufige Gemeinsamkeit der Definitionen angesehen wird. Diese Bedingung wird auch in der oft zitierten Minimaldefinition des europäischen Parlaments angeführt, nach der Künstliche Intelligenz die Fähigkeit einer Maschine sei, „menschliche Fähigkeiten wie logisches Denken, Lernen, Planen und Kreativität zu imitieren" sowie „ihr Handeln anzupassen, indem sie die Folgen früherer Aktionen analysieren und autonom arbeiten" (Europäisches Parlament 2025).

Wird im Laiendiskurs über KI diskutiert, so sind seit 2022 vor allem Large Language Modelle gemeint, die auf Basis neuronaler Netze natürliche menschliche Sprache verarbeiten und generieren können. Sie stehen auch im vorliegenden Beitrag im Mittelpunkt der Ausführungen. Sich in die grundlegende Funktionsweise dieser neuronalen Netze einzuarbeiten, erfordert eine ausführliche Auseinandersetzung mit einem spezifischen computerwissenschaftlichen Diskurs, der für viele Personen möglicherweise unzugänglich ist. Es verwundert daher nicht, dass in der Diskussion um Künstliche Intelligenz immer wieder auch Fehlvorstellungen auftauchen, wie z.B. die Beschreibung von generativer KI als Kopiermaschine bzw. als Datenbank, aus der Wissenselemente zusammenkopiert werden oder die Überzeugung, neuronale Netze würden wie klassische Computeranwendungen programmiert. Neuronale Netze zeichnen sich aber gerade dadurch aus, dass sie – anders als klassischer Computercode, der gewissen Regeln folgt und damit deterministisch ist – probabilistisch und damit non-linear operieren. Wissenselemente werden – anders als in Datenbanken – nicht an konkreten Orten gespeichert, sondern verteilen sich über verschiedene künstliche Neuronen bzw. sind nur indirekt in den sogenannten weights abgebildet. Man spricht im Fall neuronaler Netze und dem damit verbundenen deep learning vom sogenannten konnektionistischen Paradigma (vgl. Pasquinelli 2024, 24-25). Gemeint ist damit die Auffassung, dass Maschinen nur dann in die Lage versetzt werden können, komplexe menschliche kognitive Leistungen zu imitieren, wenn ihre Struktur der Funktionsweise des menschlichen Gehirns nachempfunden wird. Zentral ist, dass das daraus entstehende neuronale Netz nicht mehr in einen deterministischen Algorithmus „rückübersetzt" werden kann, wie Bajohr mit Verweis auf Offert ausführt:

> As a result, while it is technically possible to 'read' the values of the weights in a trained neural network, these numbers do not translate into a sequence of comprehensible instructions or steps in the same way that traditional code in a programming language does. (Bajohr 2024, 82)

Die Fehlvorstellung, dass ChatGPT wie eine Datenbank funktioniert, könnte nun wiederum Auswirkungen auf den Umgang mit der Technologie selbst haben: So könnte man annehmen, dass bei einer Korrektur falscher Informationen innerhalb eines Chats die notwendigen Anpassungen direkt durch eine Änderung eines Datenbankeintrags durchgeführt werden können und somit auch in Zukunft verfügbar sind. Gerade weil die verwendeten Transformer-Modelle (vgl. Vaswani et al. 2017) ihre

weights aber nicht aktualisieren (sie sind pretrained und damit fixiert nach einem Trainingsdurchgang) und in der Berechnung der Antwort – der sogenannten inference – immer wieder andere Pfade durch das komplexe Netzwerk genommen werden, ist es schwierig, das Verhalten von Large Language Modellen zuverlässig vorherzusagen. Tatsächlich kommt eine Studie in Bezug auf das technische Wissen von Lehrpersonen über neuronale Netze und LLMs zum Ergebnis, dass Wissenslücken der Lehrer:innen sowohl allgemeine technische Aspekte von KI als auch Funktionsprinzipien von KI und ihren Methoden umfassen (vgl. Lindner / Berges 2020, 8). Hierbei ist allerdings hervorzuheben, dass die befragten Personen angehende Informatiklehrkräfte waren und daher davon auszugehen wäre, dass sie einschlägiges Wissen im Bereich der Computerwissenschaft zumindest im Studium hätten erwerben können. Provokativ gefragt: Müssen computerwissenschaftliches Wissen bzw. die oben genannten technischen Grundlagen und Funktionsprinzipien generativer KI nun auch noch Bestandteil der Aus- und Weiterbildung von Deutschlehrer:innen werden? Müssen sich Fachdidaktiker:innen in computerwissenschaftliche Paper zu generativer KI einarbeiten?

2.2 HERAUSFORDERUNG WISSENSQUELLEN

Gesetzt den Fall, eine Deutschlehrkraft beschließt, sich selbstständig technisches Wissen über Large Language Modelle anzulesen, sie wäre mit einer weiteren Herausforderung konfrontiert: Wo würde sie anfangen? Welche Wissensquellen sind geeignet, um einen guten Überblick zu gewinnen?

Es ist anzunehmen, dass viele Personen sich zunächst im Internet über Large Language Modelle informieren und dabei unter Umständen sogar auf eine von generativer KI erstellte Zusammenfassung zurückgreifen (vgl. Google 2026). Informationsquellen im Internet sind mannigfaltig vorhanden und zahlreiche Beiträge versprechen eine „einfache Erklärung“ zur Funktionsweise von neuronalen Netzen und den daraus entstehenden Anwendungen (vgl. A-SIT Zentrum für sichere Informationstechnologie 2025; Mittelstand-Digital Zentrum Augsburg 2025). Darüber hinaus finden sich zahlreiche Quellen, die sich direkt an Akteur:innen im Bildungsbereich richten, wie z.B. Linksammlungen (Education Innovation Studio) oder Plattformen für Lehrende (z.B. Fobizz).

Sucht man gezielt nach Informationen in Bezug zum Deutschunterricht bzw. spezifischer zur Deutschdidaktik, kann man auf eine vielfältige Forschungsliteratur zurückgreifen: So wird in verschiedenen Themenheften (Krammer / Leichtfried 2024; Müller / Fürstenberg 2023b; Standke 2024; Rezat / Schindler 2025), Tagungen (DeutschGPT 2023; AG-Medien-Tagung 2025: DU trifft KI) und anderen Formaten wie Podcasts (Leichtfried 2024) auf das Phänomen KI reagiert. Nun wäre es im Zusammenhang dieses Beitrags von großem Interesse, diese Veröffentlichungen im Hinblick auf ihre technologisch-mediale Perspektive hin zu untersuchen und zu analysieren, wie sie die Funktionsweise von KI thematisieren bzw. wie detailliert solche Erklärungen ausfallen. Denn auf einer Metaebene stellt sich für derartige Beiträge die Herausforderung, Informationen aus einem wissenschaftlichen Feld auswerten zu müssen,

das selbst in der ausführlichen Liste der Bezugswissenschaften der Fachdidaktik Deutsch nach Abraham und Kepser (2016, 14-16) nicht vorkommt: die Computerwissenschaft.

Nachdem eine ausführliche Analyse im Rahmen dieses Beitrags nicht möglich ist, soll im Folgenden exemplarisch an zwei Beiträgen untersucht werden, ob und wie dieser Bezug zur Computerwissenschaft hergestellt wird und welche Limitationen dabei sichtbar werden. Ausgewählt wurden zwei Texte, auf die in deutschdidaktischen Forschungsbeiträgen vielfach dann verwiesen wird, wenn es um die Funktionsweise von generativer KI geht: „KI im Deutschunterricht – Funktionsprinzipien und kompetenzbezogene Einsatzmodelle" von Müller und Fürstenberg (2024) und „Der Sprachgebrauchsautomat. Die Funktionsweise von GPT und ihre Folgen für Germanistik und Deutschdidaktik" (Müller / Fürstenberg 2023a). Sie enthalten, neben Ausführungen zu den Konsequenzen der Technologie für den Deutschunterricht, jeweils Abschnitte, in denen erklärt wird, wie Transformer-Modelle (z.B. GPT) funktionieren. Anhand der folgenden kritischen Prüfung[2] und den identifizierten Limitationen stellt sich die Frage, welches Wissen für Deutschlehrkräfte „gut genug" ist und wo die Grenze zwischen (didaktisch notwendiger) Vereinfachung und Reduktionismus verläuft.

2.2.1 LIMITATION 1: ANBINDUNG AN WISSENSCHAFTLICHEN DISKURS

Es ist vor allem dem hohen Tempo technologischer und wissenschaftlicher Entwicklung und der damit verbundenen Fülle an Publikationen geschuldet, dass Unbedarfte Schwierigkeiten haben, den computerwissenschaftlichen Diskurs zu maschinellem Lernen und künstlicher Intelligenz zu überblicken. Zu prüfen wäre bei (deutschdidaktischen) Publikationen zum Thema generative KI, inwieweit Bezug auf Forschungsbeiträge aus der Computerwissenschaft genommen wird: Sind es eher zu erwartende Verweise, z.B. auf jenes vielzitierte Paper, das die Transformer-Technologie vorstellt (vgl. Vaswani et al. 2017), oder wurde der Forschungsstand ausführlich berücksichtigt? Problematisch ist eine fehlende Anbindung an den Forschungsdiskurs um generative KI insofern, als dass viele Behauptungen, wie z.B. die Aussage, dass „GPTs Sprache gar nicht verstehen können" (Fürstenberg / Müller 2024, 3) keineswegs unwidersprochen und daher Gegenstand umfassender Debatten sind (vgl. Mitchell / Krakauer 2023). Es ist demnach nicht grundsätzlich falsch, zu behaupten, Large Language Modelle hätten kein Verständnis der Welt, geboten scheint es aber, diese Behauptungen wissenschaftlich zu fundieren: So reihen sich die beiden analysierten Beiträge von Müller und Fürstenberg zwar in eine lange Tradition kritischer Analyse von Large Language Modellen, indem zentrale Kritikpunkte und sogar Begriffe wie „Papagei" aufgegriffen werden, allerdings ohne den zugrundeliegenden Beitrag „On the Dangers of Stochastic Parrots" von Bender et al. (2021) explizit zu nennen.

2.2.2 LIMITATION 2: CHERRY PICKING

Ein kontroverser Diskurs verführt darüber hinaus dazu, für die Absicherung der eigenen Position wie z.B. „Sprachmodelle können nicht verstehen!" oder „Sprachmodelle werden Lehrpersonen in nur wenigen Jahren ersetzen!" nur jene Quellen auszuwäh-

***2** Wichtig ist mir zu betonen, dass es hier dezidiert nicht um eine Kritik an den Autoren und deren Verdiensten, ein komplexes Thema zugänglich und anschaulich zu erklären, geht.*

len, die den eigenen Standpunkt bestätigen. Im Englischen hat sich dafür der Begriff „cherry picking" etabliert. Dies kann sowohl bei Lehrkräften als auch bei Wissenschaftler:innen auftreten und ist im Falle von „KI" ebenso der Komplexität des Phänomens sowie der daraus entstehenden kontroversen Diskurse geschuldet. Für „KI" als „essentially contested concept" (Gallie 1955) eine angemessene Definition zu finden, ist schwierig (vgl. Koehler 2024 oben) und durch die Vielzahl der technischen Begrifflichkeiten kommt es zu Ungenauigkeiten beim Belegen von Behauptungen.

Im analysierten Beitrag von Fürstenberg / Müller findet sich z.B. die Behauptung, GPTs verfügten über kein Modell der Welt, indem auf einen Preprint von Ha und Schmidhuber verwiesen wird. Dieser Preprint bezieht sich allerdings nicht auf die diskutierten Transformer-Modelle und zeigt eben – konträr zur Argumentation – wie in einem neuronalen Netzwerk ein internal world model aufgebaut werden kann (vgl. Ha / Schmidhuber 2018). Beiträge wie jener von Warrier et al. (2025) weisen darauf hin, dass große Sprachmodelle beim Lernen von world models im Vergleich zu Menschen zwar hinterherhinken, sie schließen allerdings auch nicht aus, dass LLMs world models lernen können.

Die Behauptung, GPTs könnten nur „den Anschein von Kommunikation, Verständnis und Argumentation bewirken" (Fürstenberg / Müller 2024, 3), wird durch einen Verweis auf Bubeck et al. 2023 belegt. Der Beleg verwundert allerdings bei näherer Betrachtung, da der Beitrag von Bubeck et al. 2023 in der computerwissenschaftlichen Community vor allem dafür kritisiert wurde, allzu optimistisch über die Fähigkeiten von GPT-4 zu berichten (vgl. Grzankowski 2024, 2-6): So argumentieren Bubeck et al. im selben Paper, das zur Untermauerung der Position dient, GPTs könnten nur den Anschein von Verständnis bewirken, dass das analysierte GPT-4 Modell sehr wohl die von Müller / Fürstenberg in Abrede gestellte Fähigkeit tiefgehenden Verständnisses zeige:

> GPT-4 demonstrates understanding of the concepts from graph theory and algorithms. […] The conversation reflects profound understanding of the undergraduate-level mathematical concepts discussed, as well as a significant extent of creativity. (Bubeck et al. 2023, 41)

Aus der großen Zahl an Publikationen im Forschungsbereich „Künstliche Intelligenz" also jene auszuwählen, die die eigene Position untermauern und davon abweichende Befunde nicht zur Kenntnis zu nehmen, deutet darauf hin, dass es generell schwierig ist, den wie auch immer gearteten aktuellen Forschungsstand ausgewogen darzustellen. Dies liegt nicht nur an der Quantität, sondern auch an dem schnellen technischen Fortschritt im Feld. So beziehen sich Studien, die beispielsweise die Limitationen von Sprachmodellen testen, oftmals auf bereits überholte Versionen (z. B. GPT-4o, GPT-o1 oder Llama 2).[3] Umso schwieriger ist es also, Quellen auszuwählen, die einen ausgewogenen Überblick über den kontroversen Diskurs geben und eben nicht nur partikuläre Positionen untermauern.

3 *Wollte man einen Überblick über die Limitationen „aktueller" Language Modelle geben, scheint sich dem Titel folgend beispielsweise das im Dezember 2024 erschienene Paper „A Primer on Large Language Models and their Limitations" von Johnson und Hyland-Wood (2024) anzubieten. Das darin getestete GPT-4o wurde allerdings am 13.02.2026 von OpenAI laut eigenem Wording „retired" (OpenAI 2026).*

2.2.3 LIMITATION 3: REDUKTIONISMUS

Eng verknüpft mit Cherry Picking ist die Herausforderung des Reduktionismus. Angesichts der Publikation der analysierten Aufsätze in Formaten, die sich einerseits an Germanist:innen (*Mitteilungen des deutschen Germanistenverbandes*) und andererseits an Deutschlehrer:innen (*Der Deutschunterricht*) richten, ist es verständlich, dass nicht auf alle Facetten der wissenschaftlichen Auseinandersetzung mit Large Language Modellen eingegangen werden kann (alleine schon wegen Restriktionen in der Anzahl der maximal zu nennenden Quellen). Veröffentlicht in reichweitenstarken Medien tragen die Beiträge allerdings zur Wissens- und Haltungsbildung einer Community bei und werden auch als state-of-the-art in anderen wissenschaftlichen Beiträgen oder studentischen Arbeiten zitiert. Reduktionistische Positionierungen in Bezug auf die Funktionsweise können dann allerdings Vorurteile in einer Community bzw. ablehnende Haltungen verstärken.[4] Eine besonders häufige Reduktion in Bezug auf generative KI ist die Aussage, Large Language Modelle würden durch Mustererkennung nur Wörter vorhersagen und zwar auf Basis einer „relativ simplen Wahrscheinlichkeitsberechnung" (Fürstenberg / Müller 2024, 2). Dass dies keineswegs unwidersprochen ist, zeigt beispielsweise die Argumentation von Boria: Zwar sei damit das LLMs zugrundeliegende Funktionsprinzip benannt, dieses als die Essenz zu setzen und die daraus emergierenden Fähigkeiten zu ignorieren, sei so, als würde man Menschen als „nur Gen-Replikations-Maschinen" bezeichnen:

> The objective that drives a system's formation – whether it's evolutionary fitness or next-token prediction – does not necessarily limit or fully describe the internal mechanics and emergent complexities that arise in pursuit of that objective. LLMs, much like our own brains, can generate rich, nuanced internal representations that go far beyond the simplistic idea of ‚predicting the next token.' (Boria 2025)

Diese nicht-reduktionistische Sichtweise wird nicht nur in wissenschaftlichen Beiträgen (Grzankowski / Downes / Forber 2025) und Blogs vertreten (Law 2025a), sondern es gibt zunehmend empirische Hinweise aus Forschungen im Bereich der „Mechanistic Interpretability" (Kästner / Crook 2024), dass es zu Phänomenen in Bezug auf neuronale Netzwerke kommt, die reduktionistische Positionen fragwürdig machen. So berichten Wissenschaftler:innen aus der Forschungsabteilung des KI-Anbieters Anthropic, dass das LLM „Claude 3.5 Haiku" elaborierte Strategien wie Vorwärts- und Rückwärtsplanen zeigt, um Aufgaben wie Übersetzung oder Wissensabfragen zu bearbeiten:

> We see signs of primitive "metacognitive" circuits that allow the model to know the extent of its own knowledge. More broadly, the model's internal computations are highly abstract and generalize across disparate contexts. (Lindsey et al. 2025)

In Bezug auf Sprache erklären die Forscher:innen, dass es in der Analyse multilingualer Schaltkreise Hinweise gebe, dass das Sprachmodell „some genuinely language-agnostic mechanisms" zeige, was nahelegen würde, dass es Konzepte „to a com-

4 So geben im Schulbarometer 2025 der Robert Bosch Stiftung 31% der Lehrkräfte an, KI nie zu nutzen, 24% nutzen KI seltener als einmal im Monat (vgl. Robert Bosch Stiftung 2025, 33).

mon ‚universal mental language' in its intermediate activations" übersetzen würde (Lindsey et al. 2025). Ein Preprint von Jha et al. (2025) kommt zu ähnlichen Ergebnissen, in Versuchen wurden beispielsweise Hinweise auf eine universelle Geometrie von embeddings gefunden. Diese Ergebnisse zeigen, dass die Forschung zu LLMs gerade erst am Anfang steht und reduktionistische Behauptungen folglich mit Vorsicht betrachtet werden sollten. Beckmann und Queloz, die aktuelle Befunde aus der mechanistic-interpretability-Forschung auswerten und darauf basierend ein Framework für das Maß des Verstehens durch generative KI entwickeln, schließen daher, dass das deflationäre Bild, LLMs würden ausschließlich statistische Mustererkennung vornehmen und kein Verstehen entwickeln, „increasingly untenable" (2025, 2) sei.[5] Die hier angeführten Arbeiten sollen nicht als Beweis für die Fehlerhaftigkeit bestimmter Positionen gewertet werden, sondern zeigen, dass die Frage nach Verstehen, Repräsentation und Weltmodellen gegenwärtig offen und kontrovers ist.

Die hier in der exemplarischen Analyse zweier deutschdidaktischer Beiträge zur Funktionsweise von generativer KI aufgeworfenen Limitationen weisen darauf hin, dass das Problem der Informationsbeschaffung nicht nur für Lehrpersonen und ihre Wissensquellen gilt, sondern auch Fachdidaktiker:innen, die zum Thema KI arbeiten, vor Herausforderungen stellt – nicht zuletzt aufgrund der rapiden Entwicklungen im Feld: Welche Quellen entsprechen dem aktuellen Stand der (computerwissenschaftlichen) Forschung? Welche Quellen sind in einen fundierten computerwissenschaftlichen Diskurs eingebettet? Ist eine Einbettung überhaupt notwendig, wenn es um ganz andere Fragestellungen geht? Hier wäre es Aufgabe, einer wissenschaftlichen Community, derlei Fragen zu diskutieren und Ansätze zu entwickeln, die Orientierung geben.

2.3 HERAUSFORDERUNG BLACKBOX

Die zuvor beschriebenen Herausforderungen eines kaum zu überblickenden wissenschaftlichen Diskurses und der damit verbundenen Unübersichtlichkeit von damit verbundenen Wissensquellen münden schließlich in der generellen Herausforderung, Large Language Modelle in Hinblick auf ihre Funktionsweise zu verstehen. Dies liegt nicht nur an der schieren Komplexität neuronaler Netzwerke, sondern auch an den opaken Strukturen der aktuell führenden (gewinnorientierten) AI-Labs, die aufgrund des Wettbewerbdrucks wenig Hinweise zu technischen Details wie Trainingsmaterial, Größe der Modelle und Finetuning[6] bzw. Alignment[7] geben. Tatsächlich ist wenig über die verwendeten Trainingsdatensätze bekannt, jene Zusammensetzung, nach der ein Großteil auf dem common crawl-Datensatz aufbaut, ist mittlerweile veraltet. Aufstellungen wie jene von Thompson (2025) können bestenfalls als Orientierung dienen, über die genaue Zusammensetzung schweigen die Unternehmen – nicht zuletzt auch aus Furcht vor Urheberrechtsklagen. Die bereits erwähnte Komplexität neuronaler Netzwerke macht es trotz Kenntnis technischer Details, wie das z.B. bei open weights-Modellen der Fall ist, schwierig, kausale Aussagen über die Funktionsweise von generativer KI zu treffen, wissen-

5 So verweisen die Forscher beispielsweise auf das noch ungeklärte Phänomen des „grokking", bei dem ein Modell nach einer langen Phase scheinbarer bloßer Memorierung der Trainingsdaten plötzlich zu einer starken Generalisierung auf neue, unbekannte Daten übergeht und dabei interne, kompaktere und strukturiertere Repräsentationen ausbildet (vgl. Beckmann / Queloz 2025, 3)

6 Die nachträgliche Anpassung des pretrained Transformers auf Basis menschlicher Präferenzen.

7 Der Versuch, die Ausgaben generativer KI mit ethischen, moralischen und rechtlichen Maßstäben abzugleichen bzw. verletzende, diskriminierende und potentiell gefährliche Einsatzszenarien zu unterbinden.

schaftliche Erklärungen sind in vielen Bereichen ausständig (vgl. Wolfram 2023). Zwar gibt es Ansätze sogenannter mechanistic interpretability – diese können laut Kästner und Crook jedoch kein umfassendes Verständnis für die Funktionsweise liefern: „They do not tell us how exactly the internal structure of the model maps inputs to outputs, neither do they allow us to predict how a system will behave in novel contexts." (Kästner / Crook 2024, 52) Pointiert bringt es der Computerwissenschaftler Sam Bowman in einem Interview auf den Punkt: „We just don't understand what's going on here. We built it, we trained it, but we don't know what it's doing." (Bowman 2023) Auch Schmitt und Heckwolf schlagen in eine ähnliche Kerbe: Man könne zwar bestimmte Effekte von LLMs plausibilisieren, sei aber nicht in der Lage, sie mathematisch abzuleiten: „Hier stellt sich dann eine abschließende Grenze für das Verstehen und Erklären ein." (Schmitt / Heckwolf 2024, 72) Aussagen wie diese, unterstreichen, weshalb eine reduktionistische Sichtweise auf das komplexe Phänomen problematisch ist und wieso es in der Auseinandersetzung mit Large Language Modellen eine differenzierte Debatte braucht, zu der nicht zuletzt auch die Deutschdidaktik einen Beitrag leisten kann.

3 — GENERATIVE KI VERSTEHEN? DEUTSCHDIDAKTISCHE ANSÄTZE EINER HERMENEUTIK DER KI

Bisher wurden drei zentrale Herausforderungen postuliert: und zwar im Zusammenhang mit Fehlvorstellungen, Wissensquellen und den generellen Grenzen des Verstehens von generativer KI. Zwar mögen diese Problematisierungen teils abstrakt und theoretisch dicht wirken, sie führen jedoch zu realen Problemen sowohl in der wissenschaftlichen Thematisierung und Erforschung als auch in der schulischen Auseinandersetzung mit generativer KI im Deutschunterricht. Provokativ formuliert: Wenn selbst wissenschaftliche Erklärungen von künstlicher Intelligenz und ihrer Funktionsweise an die Grenzen des Verstehens stoßen, wie können dann Fachdidaktiker:innen, Lehrpersonen und Schüler:innen ein tragfähiges Verständnis generativer KI und ihrer Funktionsweise entwickeln? Zwar könnte man einwenden, dass es nicht die Aufgabe von Deutschdidaktiker:innen oder Deutschlehrkräften ist, die Funktionsweise von generativer KI zu erklären oder zu verstehen, dafür gebe es die Computerwissenschaft und den Informatikunterricht[8]. Im Fach Deutsch ginge es dann eher darum, wie generative KI in den Domänen des Deutschunterrichts eingesetzt wird und welche Effekte sie auf das sprachliche, literarische und mediale Lernen von Schüler:innen hat: Lernen mit KI statt Lernen über KI (vgl. Falck 2025). In dieser Argumentation würde es sowohl für Lehrkräfte als auch für Fachdidaktiker:innen reichen, auf Laiendiskurse zurückzugreifen bzw. sich auf die oben erwähnten mitunter fraglichen Wissensquellen zu beziehen. Dieser Position sollen nun drei Ansätze entgegengehalten werden, die für eine ausführliche und z.T. auch computerwissenschaftliche Beschäftigung mit künstlicher Intelligenz plädieren, diese aber stets in Bezug zu den Feldern und Bezugswissenschaften der Deutschdidaktik setzen.

***8** Tatsächlich gibt es in Österreich Pläne, die Bezeichnung für das Fach Informatik um den Zusatz „und KI" zu erweitern.*

3.1 INTERDISZIPLINARITÄT

Wie oben erwähnt, fehlt die Computerwissenschaft in der Liste der Bezugswissenschaften der Deutschdidaktik nach Kepser und Abraham, es scheint allerdings angesichts der Ubiquität der Technologie Künstliche Intelligenz und ihres Einsatzes im schulischen Deutschunterricht geboten, zumindest über ihre Aufnahme in die Liste zu diskutieren. Wie in der Problematisierung oben offenbar wurde, sollte eine den Kriterien der Wissenschaftlichkeit genügende Auseinandersetzung mit künstlicher Intelligenz vor dem Hintergrund aktueller computerwissenschaftlicher Forschung stattfinden. Dies wäre als Reaktion auf die Wissensquellen-Problematik eine Qualitätssicherung gegen Reduktionismus und Cherry Picking. Eine derartige Interdisziplinarität, die nicht unbedingt auf die Computerwissenschaft limitiert sein muss, löst die Blackbox-Problematik zwar nicht, sie verhindert aber grobe Fehlrahmungen und dient als epistemische Sicherung. Interdisziplinarität in diesem Sinne ersetzt dann zwar keinen Konsens, ermöglicht aber bessere Begründungen unter der Bedingung von Dissens. Die Berührungsängste mögen hier für viele ungleich größer sein als bei Theater-, Literatur-, Film- oder Erziehungswissenschaft, zumindest ein fundierter Überblick des aktuellen computerwissenschaftlichen Forschungsstandes sollte allerdings Voraussetzung für Fachdidaktiker:innen sein, die zum Thema forschen oder arbeiten. Für Lehrpersonen lässt sich dies wohl kaum in diesem Umfang fordern, umso wichtiger ist es aber, in fachdidaktischer Aus- und Weiterbildung, in Publikationen und Handreichungen nicht hinter den aktuellen Kenntnisstand im Forschungsbereich Künstlicher Intelligenz zurückzufallen. Im besten Fall können hier Ressourcen der interdisziplinären Kooperation aktiviert werden, sei es in Forschungsprojekten zu KI, die Forscher:innen verschiedener Disziplinen vereinen oder schulische fächerintegrierende Projekte, die das Phänomen Künstliche Intelligenz aus mehreren fachlichen Perspektiven betrachten.

3.2 ERWEITERTES VERSTEHEN DURCH EINE HERMENEUTIK DER KÜNSTLICHEN INTELLIGENZ IM GEISTE DER KULTURWISSENSCHAFT

Eine profunde Kenntnis der Forschung zu künstlicher Intelligenz ist darüber hinaus Voraussetzung dafür, dass sich die Deutschdidaktik und ihre Vertreter:innen in der Deutung und einem erweiterten Verstehen der Technologie nicht an den Rand drängen lassen bzw. die Definition von Verstehen nicht beim Wissen über die Funktionsweise endet.[9] Ein erweitertes Verstehen meint hier (1) Kenntnis grundlegender Funktionsprinzipien, (2) Kenntnis zentraler Grenzen/Fehlermodi und (3) interpretative bzw. kulturwissenschaftliche Einordnung des Phänomens Künstliche Intelligenz. Gerade für die letztgenannte Dimension ist die kulturwissenschaftliche Fundierung der Deutschdidaktik als „eingreifende Kulturwissenschaft“ (Kepser / Abraham 2016) Anlass, selbstbewusst und von dem Verständnis der genuin germanistischen bzw. deutschdidaktischen Gegenstände bzw. Bereiche her, das Phänomen generativer künstlicher Intelligenz zu deuten und eine technizistische Sichtweise zu erweitern bzw. zu ergänzen. Der Grund dafür liegt nicht nur in der Tatsache, dass wissenschaftliche Disziplinen wie die Sprach-, Literatur- und Medienwissenschaft umfassende Expertise in Bezug auf Sprache, Symbolsysteme und ihre Phänomene haben, sondern

9 Dies gilt auch für die schulische Beschäftigung mit Künstlicher Intelligenz. Auch hier scheint der Deutschunterricht ein wichtiges Gegengewicht zu einem rein technizistischen Verständnis künstlicher Intelligenz zu sein.

dass jene Domänen auch Theorien hervorbringen und hervorgebracht haben, die das Nicht-Verstehen und die Deutungsoffenheit in besonderem Maße in den Blick nehmen (vgl. Härle 2010; Heimböckel / Pavlik 2022). Eine primär datengetriebene Erforschung künstlicher Intelligenz im Sinne der End-of-Theory-Debatte[10] würde um Ansätze erweitert, die neuronale Netzwerke und ihre Ausformungen als kulturelle Artefakte begreifen, die gelesen und interpretiert werden müssen. Dieses Verständnis legt beispielsweise Gastaldi in Bezug auf Sprachtechnologien nahe, wenn er argumentiert, dass die LLMs zugrundeliegenden embeddings – also jene vieldimensionalen Vektorräume, die die intrikaten und mannigfaltigen syntaktischen, aber eben auch semantischen Relationen zwischen Wörtern repräsentieren – nicht nur eine nützliche Technologie seien, sondern auch ein gesamtes Bild der Sprache enthalten würden, „or in other terms a general but elementary conception of what language is, of what it means to speak and write" (Gastaldi 2021, 208).

Zentral für ein kulturwissenschaftliches Verständnis von künstlicher Intelligenz ist demnach, dass die Texte, die von LLMs hervorgebracht werden, eben vor allem als Texte in den Blick geraten. Die Interpretation generativer KI und ihres Outputs könnte dann auch unter Rückgriff auf die Perspektiven und begrifflichen Instrumentarien der Literatur-, Sprach-, Medien-, und Theaterwissenschaft vorgenommen werden. So führt Blunz in Bezug auf die Literaturtheorie aus:

> Denn sich der Welt über Form anzunähern, heißt nicht, dass in dieser Art der Annäherung kein Bezug zu unserer menschlichen Realität bestünde, wie viele Romane und Kunstwerke zeigen. Und hier wird es interessant. Die künstliche Spracherzeugung entspricht sicherlich nicht unserer menschlichen Welt, aber wir können uns dennoch bemühen, die andere Art der Schrift und ihre Tendenzen zu verstehen. [...] Wenn man eine mathematisch informierte Literaturtheorie entwickeln und über Literatur durch Sprachmodelle nachdenken kann, könnten wir dann vielleicht auch den Sinn von Sprachmodellen durch Literaturtheorien erfassen? (Blunz 2024, 77-78)

Blunz entwickelt hier eine Argumentation, die anschlussfähig für den deutschdidaktischen Diskurs ist: Ein Verstehen künstlicher Intelligenz reicht über das mechanistische Verständnis der Funktionsweise hinaus. Als kulturelle Artefakte (vgl. Offert 2023) gerät generative KI in den kulturwissenschaftlichen Fokus und zwar als (noch) Unverstandenes, das zur Interpretation auffordert. Der Fachdidaktik Deutsch als integrativer Wissenschaft kommt damit in Bezug auf Künstliche Intelligenz im schulischen Kontext auch die Aufgabe zu, dieses kulturwissenschaftliche Wissen einzubringen und damit nicht zuletzt das Lernen über KI aus kulturwissenschaftlicher Sicht in der Schule mitzugestalten. Um die eingangs erwähnte Rede von dem „end of theory" aufzugreifen: Es braucht in dieser Argumentation eben gerade mehr Theorie, die nicht zuletzt auch empirische Forschungsprojekte informieren sollte. Welches Verständnis von generativer KI wird eigentlich dem Forschungsprojekt zugrunde ge-

10 *2008 sorgte ein Beitrag von Anderson mit dem Titel „The End of Theory: The Data Deluge Makes the Scientific Method Obsolete" für angeregte Debatten über die Rolle der Theorie in wissenschaftlichen Feldern, die zunehmend auf massenhafte Daten und deren Verarbeitung setzen. Anderson stellt provokant fest: „The new availability of huge amounts of data, along with the statistical tools to crunch these numbers, offers a whole new way of understanding the world. Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all. There's no reason to cling to our old ways. It's time to ask: What can science learn from Google?" Anderson (2008). Er argumentiert also, dass in einer Welt von big data, Theorien über die Welt hinfällig seien, wichtig wäre nur, was gemessen und als Effekt festgestellt werden könne.*

legt? Welche Zuschreibungen nimmt man in Bezug auf neuronale Netzwerke vor? Wie positioniert man sich in noch ungeklärten Fragen in Bezug auf LLMs? Eine Hermeneutik künstlicher Intelligenz ist somit auch Aufgabe der Deutschdidaktik und des Deutschunterrichts.[11]

3.3 INTELLEKTUELLE DEMUT UND EPISTEMISCHE OFFENHEIT

Wie oben gezeigt wurde, stellen sich für ein rein mechanistisches Verstehen künstlicher Intelligenz eine Reihe von Herausforderungen und gerade die Tatsache, dass es noch immer eine abschließende Grenze für das Verstehen und Erklären gibt, legt nahe, in Hinblick auf die Limitationen und Potentiale von Künstlicher Intelligenz – insbesondere auch für den Deutschunterricht – offen zu bleiben und reduktionistische Haltungen zu vermeiden. Solche reduktionistischen Positionen in Bezug auf generative KI, die mitunter durch die genannten Fehlvorstellungen oder fragliche Wissensquellen entstehen können, sind beispielsweise die Charakterisierung großer Sprachmodelle als „dumm", reine Kopiermaschinen oder stochastische Papageien auf der einen Seite und Heilsversprechungen für alle didaktischen Probleme durch KI auf der anderen Seite. Die Komplexität des Phänomens Künstliche Intelligenz verlangt eher eine Haltung intellektueller Demut bei gleichzeitiger Neugierde und Offenheit für neue Entwicklungen (vgl. Law 2025b). Um Beliebigkeit und Relativismus keinen Vorschub zu leisten, würde sich in rigorosen Prüfpraktiken widerspiegeln (z.B. systematisches Gegenchecken, Hypothesentesten am System, gegenteilige Argumente rezipieren). Dies gilt nicht zuletzt auch für Lehrpersonen, die dem Einsatz künstlicher Intelligenz ablehnend gegenüberstehen und/oder mit pauschalen Verboten das Thema kategorisch aus dem Unterricht ausschließen. Statt Ablehnung sollte dann eher eine dialektische Auseinandersetzung stehen, die jenseits der Binarität von Technikskepsis und Hype „dritte Denkungsarten" etabliert, wie Nassehi dies nahelegt:

> Will man intellektuell irgendetwas bewegen, muss man aus dieser Binarität ausbrechen und dritte Denkungsarten etablieren. Denn wenn es einen Beitrag von Intellektuellen, von Wissenschaftlern, von Reflexionsarbeit gibt, dann ist es sicher nicht, für operative Lösungen zu sorgen, sondern den Diskurs mit Abweichungen zu versorgen, mit mehrwertigen Beobachtungsformen, mit anderen Unterscheidungen, mit Transzendierungen binärer Denkgefängnisse. Dass es gerade nicht wenige, wenn man so will, Geistesarbeiter sind, die das kaum beherzigen, ist enttäuschend. (Nassehi 2024)

Denn liegt nicht letztlich auch hierin die Stärke jener an Sprach-, Literatur- und Medienwissenschaft geschulten Professionellen: Dass sie um die Unabschließbarkeit der Sinnbildung wissen und sie diese Offenheit gegenüber dem Unverstandenen daran erinnert, dass es ein endgültiges Wort (vgl. Härle / Steinbrenner 2010) über Texte – seien sie von Menschenhand geschrieben oder maschinengeneriert – nicht gibt.

11 *Siehe Leichtfried (2026) für eine Skizze einer derartigen Hermeneutik.*

LITERATUR

— **Anderson, Chris (2008)**: The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. In: *WIRED* vom 23.06.2008. https://www.wired.com/2008/06/pb-theory/ [20.06.2024]. — **A-SIT Zentrum für sichere Informationstechnologie (2025)**: KI-Basiswissen: Grundlagen der KI einfach erklärt. https://www.onlinesicherheit.gv.at/Services/News/KI-Grundlagen-Algorithmus.html [30.06.2025]. — **Bajohr, Hannes (2024)**: Operative ekphrasis: the collapse of the text/image distinction in multimodal AI. In: *Word & Image* 40, H. 2, 77-90. — **Beckmann, Pierre / Queloz, Matthieu (2025)**: Mechanistic Indicators of Understanding in Large Language Models. https://arxiv.org/pdf/2507.08017 [20.02.2026]. — **Bender, Emily M. et al. (2021)**: On the Dangers of Stochastic Parrots. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. New York: Association for Computing Machinery, 610-623. — **Birk, Christian (2023)**: ChatGPT im Deutschunterricht. In: *Schulmagazin* 5-10 91, 5+6, 34-39. — **Blunz, Mercedes (2024)**: Nachdenken über generatives Schreiben. In: Hannes Bajohr/Moritz Hiller (Hg.)*: Das Subjekt des Schreibens. Über Große Sprachmodelle.* München: Richard Boorberg Verlag, 71-80. — **Boria, Alejandro Tlaie (2025)**: My objection(s) to the "LLMs are just next-token predictors" take. https://alejandrotlaie.net/my-objections-to-the-llms-are-just-next-token-predictors-take [01.07.2025]. — **Bowman, Sam (2023)**: *Even the scientists who build AI can't tell you how it works.* https://www.vox.com/unexplainable/2023/7/15/23793840/chat-gpt-ai-science-mystery-unexplainable-podcast [24.03.2025]. — **Bubeck, Sébastien et al. (2023)**: Sparks of Artificial General Intelligence: Early experiments with GPT-4. https://doi.org/10.48550/arXiv.2303.12712 [16.02.2026]. — **Engelmayr-Hofmann, Katrin / Leichtfried, Matthias (2026)**: Like a Slot Machine: AI, Representation and the Challenges of Language Diversity. In: Wissia Fiorucci/Alvise Sforza Tarabochia (Hg.): *The AI Revolution in Language Education. Practices, Challenges, Future.* Palgrave Macmillan. — **Europäisches Parlament (2025)**: Was ist künstliche Intelligenz und wie wird sie genutzt? | Themen | Europäisches Parlament. https://www.europarl.europa.eu/topics/de/article/20200827STO85804/was-ist-kunstliche-intelligenz-und-wie-wird-sie-genutzt [25.06.2025]. — **Falck, Joscha (2025)**: Lernen und Künstliche Intelligenz vom 13.12.2025. https://joschafalck.de/lernen-und-ki/ [05.02.2026]. — **Fürstenberg, Maurice / Müller, Hans-Georg (2024)**: KI im Deutschunterricht. Funktionsprinzipien und kompetenzbezogene Einsatzmodelle. In: *Der Deutschunterricht* 2024, H. 5, 2-13. — **Gallie, Walter Bryce (1955)**: Essentially Contested Concepts. In: *Proceedings of the Aristotelian Society* 56, 167-198. Auch online: http://www-jstor-org.uaccess.univie.ac.at/stable/4544562. — **Gastaldi, Juan Luis (2021)**: Why Can Computers Understand Natural Language? In: *Philosophy & Technology* 34, H. 1, 149-214. — **Gesellschaft für Informatik (2016)**: Dagstuhl-Erklärung. Bildung in der digital vernetzten Welt. https://dagstuhl.gi.de/dagstuhl-erklaerung [24.05.2023]. — **Google (2026)**: Google AI Overviews - Search anything, effortlessly. https://search.google/ways-to-search/ai-overviews/ [04.02.2026]. — **Grzankowski, Alex (2024)**: Real sparks of artificial intelligence and the importance of inner interpretability. In: *Inquiry*, 1-27. — **Grzankowski, Alex / Downes, Stephen M. / Forber, Partick (2025)**: LLMs are not just next token predictors. In: *Inquiry*, 1-11. — **Ha, David / Schmidhuber, Jürgen (2018)**: Recurrent World Models Facilitate Policy Evolution. http://arxiv.org/pdf/1809.01999 [16.02.2026]. — **Härle, Gerhard (2010)**: Irritation und Nicht-Verstehen. Zur Hermeneutik als Provokation für die Literaturdidaktik. In: Michael Baum/Marion Bönnighausen (Hg.): *Kulturtheoretische Kontexte für die Literaturdidaktik*. Baltmannsweiler: Schneider Verlag Hohengehren, 10-23. — **Härle, Gerhard / Steinbrenner, Marcus (Hg.) (2010)**: *Kein endgültiges Wort. Die Wiederentdeckung des Gesprächs im Literaturunterricht*. 2. Aufl. Baltmannsweiler: Schneider Verlag Hohengehren. — **Heimböckel, Hendrick / Pavlik, Jennifer (Hg.) (2022)**: *Ästhetisches Verstehen und Nichtverstehen. Aktuelle Zugänge in Literatur- und Mediendidaktik.* Bielefeld: transcript Verlag. (= Lettre, Bd. 2). — **Hesse, Florian / Jagdschian, Larissa Carolin (Hg.) (2026, im Erscheinen)**: *Literatur \\ Künstliche Intelligenz // Didaktik*. Würzburg: Königshausen & Neumann. — **Jha, Rishi et al. (2025)**: Harnessing the Universal Geometry of Embeddings. 10.48550/arXiv.2505.12540 [16.02.2026]. — **Johnson, Sandra / Hyland-Wood, David (2024)**: A Primer on Large Language Models and their Limitations. https://arxiv.org/pdf/2412.04503 [18.06.2026]. — **Kästner, Lena / Crook, Barnaby (2024)**: Explaining AI through mechanistic interpretability. In: *European Journal for Philosophy of Science* 14, H. 4, 51-76. — **Kepser, Matthis / Abraham, Ulf (2016)**: *Literaturdidaktik Deutsch. Eine Einführung*. 4. Aufl. Berlin: Erich Schmidt Verlag. — **Koehler, Jana (2024)**: Zum Begriff der ‚Künstlichen Intelligenz'. In: Stephanie Catani (Hg.)*: Handbuch Künstliche Intelligenz und die Künste.* Berlin: De Gruyter, 9-26. — **Krammer, Stefan / Leichtfried, Matthias (Hg.) (2024)**: *Künstliche Intelligenz. Auswirkungen und Anwendungen im Deutschunterricht.* Innsbruck: Studien Verlag. — **Law, Harry (2025a)**: Yes, linear algebra can ‚know' things. https://www.learningfromexamples.com/p/does-ai-know-things [01.07.2025]. — **Law, Harry (2025b)**: Academics are kidding themselves about AI. https://www.learningfromexamples.com/p/what-academics-get-wrong [01.07.2025]. — **Leichtfried, Matthias (2024)**: KI im Deutschunterricht - Perspektiven auf Theorie und Praxis. https://open.spotify.com/show/0HI6PwafjYYv3ZuMeChFxy [16.02.2026]. — **Leichtfried, Matthias (2026):** Grenzgänge. Skizze einer Hermeneutik Künstlicher Intelligenz als Bezugspunkt literaturdidaktischer Auseinandersetzung. In: Florian Hesse/Larissa Carolin Jagdschian (Hg.): Literatur - Künstliche Intelligenz - Didaktik. Würzburg: Königshausen & Neumann, 179-199. — **Lindner, Annabel / Berges, Marc (2020)**: Can you explain AI to me? Teachers' pre-concepts about Artificial Intelligence. In: *Education for a sustainable future.* IEEE, 1-9. — **Lindsey, Jack et al. (2025)**: On the Biology of a Large Language Model. In: *Transformer Circuits Thread*. Auch online: https://transformer-circuits.pub/2025/attribution-graphs/biology.html [16.02.2026]. — **Long, Duri / Magerko, Brian (2020)**: What is AI Literacy? Competencies and Design Considerations. In: Regina Bernhaupt (Hg.): *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* New York: Association for Computing Machinery, 1-16. — **Mitchell, Melanie / Krakauer, David C. (2023)**: The

debate over understanding in AI's large language models. In: *Proceedings of the National Academy of Sciences of the United States of America* 120, H. 13, 1-5. — **Mittelstand-Digital Zentrum Augsburg (2025)**: Was ist Künstliche Intelligenz? KI einfach erklärt - Mittelstand-Digital. https://digitalzentrum-augsburg.de/kuenstliche-intelligenz-einfach-erklaert/ [30.06.2025]. — **Müller, Hans-Georg / Fürstenberg, Maurice (2023a)**: Der Sprachgebrauchsautomat. Die Funktionsweise von GPT und ihre Folgen für Germanistik und Deutschdidaktik. In: *Mitteilungen des Deutschen Germanistenverbandes* 70, H. 4, 327-345. — **Müller, Hans-Georg / Fürstenberg, Maurice (Hg.) (2023b)**: *Digitale Umbrüche. Sprache, Literatur und Deutschunterricht in Zeiten von Big Data und KI.* (= Bd. 70). — **Nassehi, Armin (2019)**: *Muster. Theorie der digitalen Gesellschaft*. München: C.H. Beck. — **Nassehi, Armin (2024)**: Montagsblock /273. https://kursbuch.online/montagsblock-273/?utm_source=mailpoet&utm_medium=email&utm_source_platform=mailpoet&utm_campaign=montagsblock-113-kursbuch-27-07-2020_43 [13.05.2024]. — **Offert, Fabian (2023)**: Can We Read Neural Networks? Epistemic Implications of Two Historical Computer Science Papers. In: *American Literature* 95, H. 2, 423-428. — **OpenAI (2026):** Retiring GPT-4o, GPT-4.1, GPT-4.1 mini, and OpenAI o4-mini in ChatGPT. https://openai.com/index/retiring-gpt-4o-and-older-models/ [13.02.2026]. — **Pasquinelli, Matteo (2024)**: *Das Auge des Meisters. Eine Sozialgeschichte Künstlicher Intelligenz.* 1. Aufl. Münster: UNRAST. — **Peeperkorn, Max et al. (2024)**: Is Temperature the Creativity Parameter of Large Language Models? In: Kazjon Grace et al. (Hg.): *Proceedings of the 15th International Conference on Computational Creativity, ICCC 2024, Jönköping, Sweden, June 17-21*, 2024. Association for Computational Creativity (ACC), 226-235. — **Reusser, Kurt / Pauli, Christine (2014)**: Berufsbezogene Überzeugungen von Lehrerinnen und Lehrern. In: Ewald Terhart (Hg.): *Handbuch der Forschung zum Lehrerberuf.* 2. Aufl. Münster: Waxmann, 642-661. — **Rezat, Sara / Schindler, Kirsten (Hg.) (2025)**: *KI und Schreiben.* Friedrich Verlag. (= Praxis Deutsch, Bd. 52). — **Robert Bosch Stiftung (2025):** Deutsches Schulbarometer: Befragung Lehrkräfte. Ergebnisse zur aktuellen Lage an allgemein- und berufsbildenden Schulen. Stuttgart. — **Schmitt, Marco / Heckwolf, Christoph (2024)**: KI zwischen Blackbox und Transparenz. Das Koppeln und Entkoppeln von Kontrollprojekten. In: Roger Häußling/Claudius Härpfer/Marco Schmitt (Hg.): *Soziologie der Künstlichen Intelligenz*. Bielefeld, transcript Verlag, 51-83. — **Standke, Jan (Hg.) (2024)**: *Künstliche Intelligenz. Verhandlungen in der Gegenwartsliteratur und Perspektiven für das literarische Lernen.* Trier: wvt. — **Steinhoff, Torsten (2023)**: Der Computer schreibt (mit). Digitales Schreiben mit Word, Whatsapp, ChatGPT & Co. als Koaktivität von Mensch und Maschine. In: *MiDU – Medien im Deutschunterricht* H. 1, 1-16. https://doi.org/10.18716/ojs/midu/2023.1.4 — **Thompson, Alan D. (2026)**: The Memo - 14/Feb/2026. https://substack.com/home/post/p-187255480 [16.02.2026] — **Vaswani, Ashish et al. (2017)**: Attention Is All You Need, 1-15. Auch online: https://arxiv.org/pdf/1706.03762. — **Warrier, Archana et al. (2025)**: Benchmarking World-Model Learning. https://arxiv.org/pdf/2510.19788 [16.02.2026]. — **Wolfram, Stephen (2023)**: What Is ChatGPT Doing … and Why Does It Work? https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/ [16.02.2026].

ÜBER DEN AUTOR

Dr. Matthias Leichtfried ist Post-Doc am Institut für Germanistik der Universität Wien. Er forscht zur Literatur- und Mediendidaktik, insbesondere zu KI und Deutschunterricht in einer Kultur der Digitalität. Er ist neben seiner wissenschaftlichen Tätigkeit in der Lehrer:innenbildung aktiv, zuvor war er Lehrer für Deutsch und Philosophie/Psychologie an einem Wiener Gymnasium.