SPRACHMODELLE

Marktlage, Herausforderungen und ein Praxisvergleich aus der Biologie

Seit dem ersten Hype um ChatGPT haben sich large language models ***(LLMs; dt. große Sprachmodelle; siehe Kasten) rasant weiterentwickelt. Diese Dynamik hat nicht nur die technologische Landschaft, sondern auch die strategische Ausrichtung zahlreicher Unternehmen verändert.***

Wie stark diese Entwicklung den Markt bereits prägt, lässt sich an aktuellen Investitionszahlen eindrucksvoll ablesen. Letztes Jahr stiegen die Ausgaben für Anwendungen der Künstlichen Intelligenz (KI) gegenüber dem Vorjahr um das Sechsfache [1]. In Deutschland investierten 2024 bereits 37 Prozent der Unternehmen in KI-Technologien, während 74 Prozent angaben, dies künftig tun zu wollen [2]. Der Blick auf die erwarteten Effekte ist noch zwiegespalten: 55 Prozent der Unternehmen erhoffen sich einen Produktivitätsgewinn, gleichzeitig äußern jedoch 37 Prozent Bedenken hinsichtlich Arbeitsplatzverlusten, Datenschutzfragen und fehlender Akzeptanz in der Belegschaft [3]. Besonders weit fortgeschritten ist die Entwicklung im Gesundheitssektor: Laut der Investmentbank *Morgan Stanley* setzen bereits 94 Prozent der befragten Unternehmen in diesem Bereich KI ein [4]. Parallel dazu legten 2024 die Investitionen im Bereich *digital health* europaweit um 19 Prozent auf 3,5 Milliarden US-Dollar zu. Deutschland belegte mit 739 Millionen US-Dollar den zweiten Platz hinter dem Vereinigten Königreich (959 Millionen US-Dollar). Erwähnenswert ist, dass KI-gestützte Gesundheitsunternehmen mit 61 Prozent den größten Teil dieser Investitionen ausmachen [5].

In der Forschung zeichnet sich währenddessen eine neue Qualität der Anwendungsmöglichkeiten ab. Spezialisierte medizinische LLMs wie *Googles* MedGemma analysieren klinische Bilder und Befunde in Echtzeit direkt vor Ort, ermöglicht durch geringere Hardwareanforderungen [6]. Parallel dazu entstehen biologische Sprachmodelle, die anstelle menschlicher Sprache mit DNA-Sequenzen trainiert werden. Ein Beispiel ist Evo-2 von *Arc Institute*, das genomische Sequenzen interpretieren und neu entwerfen kann [7]. Aktuell entstehen sogar Modelle, die die Analyse biologischer Sequenzen mit den klassischen Fähigkeiten großer Sprachmodelle kombinieren. Das DNA-LLM *BioReason* kann DNA-Sequenzen nicht nur verarbeiten und interpretieren, sondern auch ähnlich wie ein LLM in Textform argumentieren und Ergebnisse präsentieren [8].

Praxisvergleich – Sprachkomplexität von LLM-Texten im Fach Biologie

Eine aktuelle Studie des US-amerikanischen KI-Anbieters *Anthropic* gibt interessante Einblicke in den fachlichen Nutzungskontext ihrer Sprachmodelle der *Claude* Reihe: Obwohl nur 5,4 Prozent der US-Bachelorabschlüsse auf Informatik entfallen, machten Konversationen mit Fokus auf Programmiersprachen 38,6 Prozent der akademisch motivierten Nutzung ihrer Sprachmodelle aus. In den Naturwissenschaften lagen die entsprechenden Anteile bei 9,2 Prozent (Abschlüsse) und 15,2 Prozent (Nutzung), während medizinische Fachrichtungen eine umgekehrte Tendenz aufwiesen: Trotz 13,1 Prozent Studienabschlüssen entfielen nur 5,5 Prozent der Konversationen mit *Claude* auf diesen Bereich [9].

Wenn rund 15 Prozent der Nutzung der LLMs bei *Anthropic* auf Aufgaben im Bereich der Naturwissenschaften zurückzuführen sind, interessiert Sie nun vielleicht, wie sich eigentlich die sprachliche Qualität LLM-generierter deutschsprachiger Texte in unserem Fachbereich der Biologie einordnen lässt. In einem kleinen, explorativen Vergleich für diesen Artikel habe ich vier der beliebtesten, kostenlos verfügbaren LLMs (*GPT-4o* (*OpenAI*, USA), *Claude Sonnet 4* (*Anthropic*, USA), *Gemini 2.5 Pro* (*Google*, USA) und *DeepSeek-R1* (*DeepSeek*, China)) mit der simplen Eingabeaufforderung (engl. *Prompt*) „Schreibe einen Artikel über das Immunsystem der Pflanzen“ auf die sprachliche Komplexität der generierten Texte untersucht. Als Referenz für von Menschen verfasste Texte wurden ein einschlägiger Artikel aus der *BiuZ* sowie ein weiterer aus einem wis-

ZUR PERSON

Bild: DSEE/bundesfoto

Arian Abbasi ist Vorstandsmitglied im Landesverband NRW des VBIO und unterstützt diesen als Bioinformatikmasterand in IT-bezogenen Themen. Beruflich ist er in der Forschung und Entwicklung im Bereich KI-Sicherheit tätig. Zudem vermittelt er Kindern und Jugendlichen bei InteGREATer e.V. den verantwortungsvollen Umgang mit Künstlicher Intelligenz.

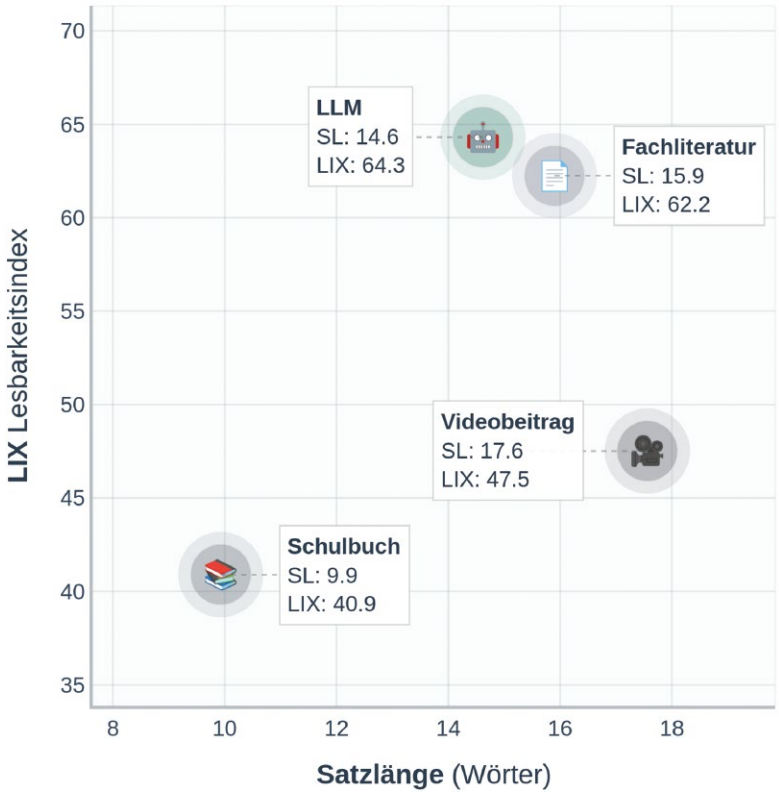

ABB. 1 Durchschnittliche Satzlänge (SL) und LIX-Lesbarkeitsindex verschiedener Textgruppen im Fachbereich Botanik. Dargestellt sind die Mittelwerte von Texten, die von vier *large language models* (LLMs) generiert wurden, im Vergleich zu den Mittelwerten zweier wissenschaftlicher Fachartikel, eines Abschnitts aus einem Biologieschulbuch der Unterstufe, sowie des Transkripts eines populärwissenschaftlichen Videobeitrags. Abb. erstellt von A. Abbasi.

senschaftlichen Journal herangezogen, die beide vor 2022 und damit vor dem breiten Einsatz großer Sprachmodelle veröffentlicht wurden. Als Kontrollgruppe, die eine geringere sprachliche Komplexität aufweisen sollte, wurden ein Abschnitt eines Biologieschulbuchs der Unterstufe sowie das Transkript eines populärwissenschaftlichen Videobeitrags zum pflanzlichen Immunsystem in die Analyse einbezogen.

Für die Bewertung der sprachlichen Qualität berechnete ich für jede Gruppe den LIX-Lesbarkeitsindex [10], eine für die deutsche Sprache geeignete Metrik aus dem Bereich der natürlichen Sprachverarbeitung (engl. *Natural Language Processing, NLP*). Der Index misst den sprachlichen Schwierigkeitsgrad eines Textes anhand der durchschnittlichen Satzlänge und des Anteils komplexer Wörter. Die so ermittelten Werte stelle ich jeweils der durchschnittlichen Satzlänge (SL) gegenüber.

Das Ergebnis (Abbildung 1): Der Mittelwert der von den vier LLMs erzeugten Texte zeigte mit einer durchschnittlichen Satzlänge von 14,6 Wörtern und einem LIX-Wert von 64,3 eine ähnliche sprachliche Komplexität wie die Fachartikel (SL: 15,9; LIX: 62,2). Zum Vergleich: Der Schulbuchtext lag wie erwartet deutlich darunter (SL: 9,9; LIX: 40,9), während der transkribierte Vortrag mit längeren Sätzen (SL: 17,6) und moderater Komplexität (LIX: 47,5) eine Zwischenposition einnahm. Die Werte lassen also vermuten, dass LLMs bereits ohne besonders detaillierte Anweisungen sprachlich eine Qualität erreichen können, die der wissenschaftlichen Fachliteratur nahekommt. Der Lesbarkeitsindex ist jedoch nur eine der vielen NLP-Methoden, und berücksichtigt nicht die inhaltliche Qualität von Texten. Bei der Nutzung von LLMs ist es ratsam, deren spezifische Eigenschaften und Einschränkungen im Blick zu behalten.

Zwischen Effizienzgewinn und Denkfaulheit

Ein grundlegendes Problem, das den Einsatz von Sprachmodellen in For-

GROSSE SPRACHMODELLE

Large language models (LLMs; dt. große Sprachmodelle) sind leistungsfähige künstliche neuronale Netzwerke mit häufig mehreren Milliarden einstellbaren Parametern. Sie werden mit umfangreichen Textkorpora aus Büchern und dem Internet trainiert. Dadurch sind sie in der Lage, komplexe sprachliche Muster und Zusammenhänge zu erkennen, Sprache zu verstehen und eigenständig zu generieren. LLMs finden zunehmend Anwendung in Forschung und Industrie. Sie können fachübergreifend Texte analysieren, strukturierte Antworten erzeugen und eine Vielzahl sprachbezogener Aufgaben übernehmen, etwa in der medizinischen Dokumentation, der juristischen Textanalyse oder im automatisierten Kundensupport. Neben Modellen, die Bilder, Videos oder Ton generieren können, bilden sie die Hauptmodalitäten der Generativen Künstlichen Intelligenz ab.

schung und Industrie begleitet, sind Fehlinformationen in den von LLMs erstellten Texten, die zwar sprachlich einwandfrei wirken, aber inhaltlich falsch sind. Im Fachjargon werden sie als „Halluzinationen" bezeichnet. Diese werden durch fehlende, fehlerhafte oder verzerrte Trainingsdaten begünstigt. Besonders in sensiblen Bereichen wie der Medizin kann das hochproblematisch werden [11, 12]. Ebenso gefährlich sind sogenannte *Prompt-Injection*-Angriffe (angelehnt an *SQL-Injections*, einen klassischen Cybersecurity-Angriff auf Datenbanken, die die Programmiersprache *Structured Query Language*, kurz SQL, benutzen). Dabei können Hacker mithilfe bösartiger *Prompts* Anwendungen, die LLMs benutzen, gezielt zum Absturz bringen oder dazu zwingen, vertrauliche oder gefährliche Informationen auszugeben [13].

Meine Beobachtungen aus der Bildungspraxis zeigen, dass Generative KI faszinierende Möglichkeiten zur kreativen Entfaltung bietet. Allerdings lassen Schülerinnen und Schüler häufig Aufgaben vollständig von LLMs erledigen, ohne das zugrunde liegende Problem wirklich zu verstehen oder kritisch zu hinterfragen. Dies wirft die Frage auf, wie sich dieser Komfort langfristig auf das Lernen auswirkt. Bedenklich ist eine vorläufige Elektroenzephalogramm-Studie des *MIT Media Lab*, einem Forschungslabor am *Massachusetts Institute of Technology*: Mit ChatGPT bearbeiteten Teilnehmende Aufgaben zwar 60 Prozent schneller, jedoch weisen die gemessenen Hirnaktivitäten darauf hin, dass die geringere geistige Anstrengung langfristig die Entwicklung und Erhaltung unserer kognitiven Fähigkeiten beeinträchtigen könnte [14].

„Responsible AI": Durch verantwortungsvollen Umgang mit KI den Nutzen maximieren

Trotz der angesprochenen Risiken sollten Sie sich nicht abschrecken lassen, sondern sich aktiv mit Sprachmodellen auseinandersetzen, um mit dem rasanten Entwicklungstempo Schritt zu halten. Wichtig ist hierbei, die von LLMs erstellten Inhalte stets kritisch zu prüfen und eigenständig zu bewerten, da Fehler oder Fehlinformationen nicht immer sofort erkennbar sind. Vor allem Jugendliche sollten frühzeitig über die Mechanismen und Einschränkungen von KI und LLMs aufgeklärt werden. So lernen sie, die Schwächen aktiv zu erkennen und sich mit den Herausforderungen auseinanderzusetzen, anstatt naiv der Technologie zu vertrauen. Mit diesem kritischen Bewusstsein können sie KI verantwortungsvoll nutzen und gleichzeitig von ihren vielfältigen Möglichkeiten profitieren.

Für Endnutzer, die KI-Modelle offline direkt auf ihrem eigenen Gerät nutzen möchten und nicht auf große KI-Anbieter außerhalb der Europäischen Union angewiesen sein wollen, stehen inzwischen leistungsstarke kleinere Open-Source-LLMs über Tools wie *Ollama*, *LM-Studio* (PC) oder die *Google AI Edge Gallery* (Mobiltelefon) zur Verfügung. Diese Modelle sind eine wertvolle Ergänzung für den Einsatz in Bereichen mit sensiblen Daten, z. B. im Labor oder der Arztpraxis, wo Datenschutz entscheidend ist, auch wenn sie bei komplexen Aufgaben noch nicht das Qualitätsniveau der kommerziellen LLMs erreichen.

Literatur

[1] Menlo Ventures (2024). The State of Generative AI in the Enterprise, https://menlovc.com/2024-the-state-of-generative-ai-in-the-enterprise. [abgerufen am 24.06.2025].

[2] Investment in AI by companies in Germany in 2024 [Graph], Bitkom, October 18, 2024, https://www.statista.com/statistics/1535862/ai-investment-companies-germany/

[3] KPMG (2024). German companies are planning significant investments in generative AI, https://kpmg.com/de/en/home/media/press-releases/2024/05/german-companies-plan-substantial-investments-in-generative-ki.html. [abgerufen am 24.06.2025].

[4] Morgan Stanley (2023). How Artificial Intelligence Could Reshape Health Care, https://www.morganstanley.com/ideas/ai-in-health-care-forecast-2023. [abgerufen am 24.06.2025].

[5] Galen Growth (2024). Europe Digital Health Funding Soars 19Prozent YoY as AI Ventures Thrive and Partnerships Surge in Q3 2024, https:https://www.galengrowth.com/europe-digital-health-funding-soars-19-yoy-as-ai-ventures-thrive-and-partnerships-surge-in-q3-2024/. [abgerufen 24.06.2025].

[6] Google (2025). MedGemma Hugging Face. https://huggingface.co/collections/google/medgemma-release-680aade845f90bec6a3f60c4. [abgerufen 24.06.2025].

[7] G. Brixi et al. (2025). Genome modeling and design across all domains of life with Evo 2. BioRxiv 2025-02.

[8] A. Fallahpour et al. (2025). BioReason: „Incentivizing Multimodal Biological Reasoning within a DNA-LLM Model." arXiv preprint arXiv:2505.23579.

[9] K. Handa et al. (2025). Anthropic Education Report: How University Students Use Claude, https://www.anthropic.com/news/anthropic-education-report-how-university-students-use-claude. [abgerufen 24.06.2025].

[10] C. H. Björnsson (1968). Läsbarhet. Lund: Liber.

[11] M. Omar et al. (2025). Large Language Models Are Highly Vulnerable to Adversarial Hallucination Attacks in Clinical Decision Support: A Multi-Model Assurance Analysis. medRxiv 2025-03.

[12] Y. Kim et al. (2025). Medical hallucinations in foundation models and their impact on healthcare. arXiv preprint, arXiv:2503.05777.

[13] MITRE, "AML.T0051" ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems. https://atlas.mitre.org/techniques/AML.T0051. [abgerufen 24.06.2025].

[14] N. Kosmyna et al. (2025). Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task. arXiv preprint, arXiv:2506.08872.

Arian Abbasi, Köln