Teil 7: Künstliche Intelligenz – Werkzeug, Schreibbegleitung und Sparringpartnerin

Martin Bettschart und **Marcel Herrmann**

Inhaltsübersicht

7.1 Künstliche Intelligenz

Inhaltsübersicht

Begriffsbestimmung und Funktionsweise

Künstliche Intelligenz (KI; *artificial intelligence, AI*) ist kein neues Phänomen. Die Forschung in diesem Bereich begann bereits in den 1940er-Jahren. Im Jahr 1950 formulierte der britische Mathematiker und Informatiker Alan Turing ein Vorgehen zur Feststellung maschineller Intelligenz, welches später als Turing-Test bekannt wurde (Turing, 1950). Ab den 1960er-Jahren bis in die 2000er-Jahre folgten durch technologische Fortschritte mehrere Phasen des Aufschwungs, geprägt von Euphorie, optimistischen Vorhersagen und grossen Investitionen in KI-Entwicklungsprojekte. Diese Phasen waren jeweils gefolgt von einsetzender Enttäuschung, weil die Entwicklung aus unterschiedlichen Gründen stagnierte und infolgedessen Fördergelder gestrichen wurden (teilweise auch als «KI-Winter» bezeichnet).[1]

Die letzte Phase des Aufschwungs erfuhr ihren (bisherigen) Höhepunkt im November 2022 mit der Veröffentlichung von ChatGPT durch OpenAI (Albrecht, 2023). Spätestens zu diesem Zeitpunkt wurde KI und deren Möglichkeiten einer breiten Öffentlichkeit bekannt.

1 Für ausführlichere historische Übersichten, siehe z. B. Deutscher Ethikrat (2023), Maggiori (2023) oder Russell und Norvig (2021).

> "AI is going to be debated as the hottest topic of 2023. And you know what? That's appropriate. This is every bit as important as the PC, as the internet."
>
> (KI wird als das heisseste Thema des Jahres 2023 diskutiert werden. Und wissen Sie was? Das ist berechtigt. Dies ist genauso bedeutend wie der PC, wie das Internet.)
>
> —Bill Gates in einem Interview mit Forbes am 2. Februar 2023 (Konrad & Cai, 2023)

Die Wellen, welche die Veröffentlichung von ChatGPT schlug, waren enorm. Es schien, als müsse sich jede und jeder dazu positionieren – das Thema war sprichwörtlich «in aller Munde». Die Auswirkungen auf Gesellschaft, Wirtschaft und Wissenschaft wurden heiss diskutiert (und werden es immer noch), wobei die Meinungen weit auseinander gingen. Die Bandbreite reichte von dystopischen Weltuntergangsszenarien (z. B. Center for AI Safety, 2023) bis zu utopisch anmutenden Zukunftsvisionen, bei denen sich KI in nahezu allen Lebensbereichen positiv auswirken würde (z. B. World Economic Forum, 2023).

Auch spezifisch auf das Schreiben bezogen gestalteten sich die Voten sehr divers. Einerseits war von möglichen negativen Auswirkungen von selbst schreibender KI zu lesen, wie z. B. irreversiblen Kompetenzverlusten oder sinkender Textqualität bis hin zu vollständig gehaltlosen Texten (Oertner, 2024; Reinmann, 2023). Andererseits wurde auf die neue Möglichkeit des «Schreibens auf Knopfdruck» hingewiesen. Innert kürzester Zeit wurden unterschiedlichste Werke mit ChatGPT generiert und werbewirksam veröffentlicht – u. a. verschiedene Kolumnen, Zeitschriftenaufsätze, Bücher, Reden und Predigten sowie die Folge einer Comedy-Serie (Albrecht, 2023).

AI literacy

> "AI becomes a fundamental skill for everyone, not just for computer scientists. In addition to reading, writing, arithmetic and digital skills, we should add AI to every learners' twenty-first century technological literacy in work settings and everyday life."
>
> (KI wird zu einer grundlegenden Kompetenz für alle, nicht nur für Informatiker:innen. Zusätzlich zu Lesen, Schreiben, Rechnen und digitalen Kompetenzen sollten wir KI für alle Lernenden zu den technologischen Kompetenzen im 21. Jahrhundert in der Berufswelt und im Alltag ergänzen.)
>
> —Ng et al. (2021, p. 9)

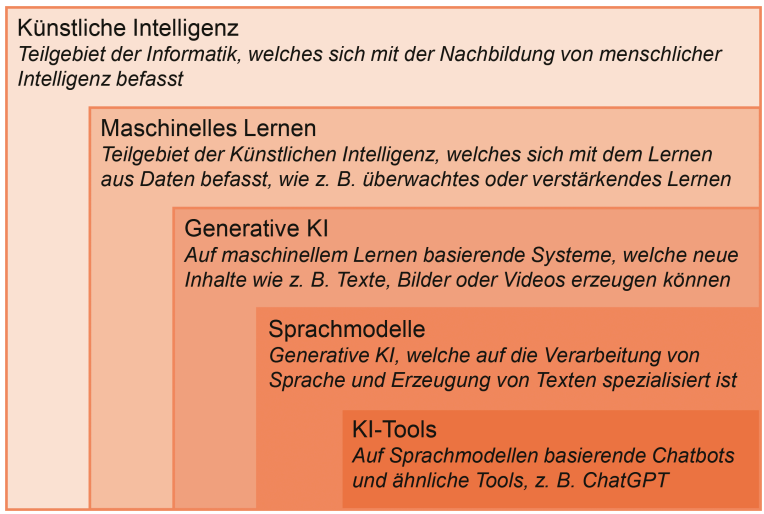
Moderate Stimmen plädieren für einen verantwortungsvollen Umgang mit KI – sowohl allgemein als auch spezifisch beim Schreiben –, bei welchem gleichzeitig den mit KI einhergehenden Risiken Rechnung getragen und die sich ergebenden Chancen möglichst gut genutzt werden. Dazu bedarf es einer neuen Fähigkeit, welche verschiedentlich als *AI literacy* bezeichnet wird. *Literacy* meint im Englischen eigentlich die Lese- und Schreibkompetenz, womit bei *AI literacy* im übertragenen Sinn von einer Kompetenz im Umgang mit KI gesprochen werden kann. Dazu gehören insbesondere (a) ein grundsätzliches Verständnis der Funktionsweise, (b) Kenntnis der ethischen und rechtlichen Rahmenbedingungen sowie (c) der sich aus der Nutzung von KI ergebenden Chancen und Risiken und (d) die Fähigkeit, KI bei Aufgaben praktisch anzuwenden (Ng et al., 2021; Southworth et al., 2023; Walter, 2024a; L. Yan et al., 2024).[2]

Diese *AI literacy* ist nicht nur für Schreibende von höchster Bedeutung, wenn sie KI nutzen. Auch für Lehrende bedeutet die Verfügbarkeit von KI für Lernende, sich mit dieser Kompetenz bzw. den didaktischen Konsequenzen der Verfügbarkeit von KI auseinandersetzen zu müssen. Lehrende sollten nicht nur selbst über *AI literacy* verfügen, sondern diese auch an Lernende weitergeben können (Filipović et al., 2025; Kultusministerkonferenz der Bundesrepublik Deutschland, 2025). Damit verbunden ist ein Umdenken in Bezug auf die Vermittlung von Kompetenzen erforderlich. Es gilt abzuschätzen, welche Kompetenzen zukünftig an Bedeutung gewinnen werden und daher vermittelt werden sollten (Reinmann, 2023). Darüber hinaus besteht die Herausforderung darin, grundlegende Kompetenzen auch angesichts der Möglichkeiten durch KI weiterhin zu sichern und gezielt um solche Fähigkeiten zu erweitern, die für das KI-Zeitalter unverzichtbar bleiben – allen voran das kritische Denken, aber auch Problemlösefähigkeit und Urteilsvermögen. Diese sog. höheren kognitiven Fähigkeiten sind nicht nur im Umgang mit KI entscheidend, sondern bilden die Grundlage, um verantwortungsvolle und komplexe Aufgaben übernehmen zu können. Werden sie vernachlässigt, drohen gravierende Konsequenzen – etwa eine wachsende Abhängigkeit von KI, ein Verlust an Selbstständigkeit im Denken sowie Fehlentscheidungen im beruflichen und gesellschaftlichen Alltag (Brommer et al., 2023; Filipović et al., 2025).

2 Es existiert aktuell keine einheitliche Definition zu *AI literacy*. Den Kern dieser Kompetenz bilden in der Regel das Verständnis der Funktionsweise (inkl. deren kritische Beurteilung) sowie die Fähigkeit zur praktischen Anwendung von KI (Laupichler et al., 2022; Long & Magerko, 2020).

Um die grundsätzliche Funktionsweise von KI zu verstehen, werden in diesem Kapitel verschiedene relevante Begriffe vorgestellt. Die Beziehungen zwischen verschiedenen relevanten Begriffen in Bezug auf das Schreiben mit KI sind in Abbildung 21 dargestellt.

Abbildung 21: Beziehungen zwischen Begriffen rund um KI



Notizen. Die Abbildung visualisiert die Beziehungen zwischen den relevantesten vorgestellten Begriffen rund um KI. Eigene Darstellung basierend auf Buck (2025) und Gimpel et al. (2023). KI-Tools basieren auf Sprachmodellen, d. h. generativen KI-Systemen, welche auf Grundlage von maschinellem Lernen – einem Teilgebiet der künstlichen Intelligenz – Sprache verarbeiten und Texte erzeugen können.

Künstliche Intelligenz

Mit dem Begriff *«künstliche Intelligenz»* (KI) sind verschiedene weitere Konzepte assoziiert, welche teils abgrenzend und teils synonym verwendet werden. KI kann dabei als Sammelbegriff verstanden werden, der alle Systeme umfasst, welche entwickelt wurden, um menschliche Intelligenz nachzubilden (Russell & Norvig, 2021). Ziel aller Entwicklungen im Bereich der KI ist die Auslagerung von Aufgaben, welche bislang von Menschen übernommen wurden, an eigens dafür entwickelte Systeme (Russell & Norvig, 2021).

KI-Systeme werden in sog. *«schwache»* und *«starke» KI* unterteilt (Searle, 1980). Bislang existieren lediglich unterschiedliche Formen von sog. schwacher KI. Damit sind Systeme gemeint, welche auf die Erledigung von spezifischen Aufgaben ausgerichtet sind. So gibt es z. B. KI-Systeme, welche gesprochene Worte in geschriebenen Text umwandeln können (Transkription). Wiederum andere Systeme, sog. *Spam-Filter,* können E-Mails auf verdächtige Inhalte durchsuchen. Auch bei KI-Tools wie z. B. ChatGPT handelt es sich gemäss Definition um schwache KI. Teilweise wird hier auch der Begriff *breite KI* verwendet, um auszudrücken, dass deren Anwendungsbereich deutlich breiter ist als bei anderer schwacher KI. Die Kategorisierung nach *enger vs. breiter KI* ist dabei als Kontinuum im Bereich der schwachen KI zu verstehen (Deutscher Ethikrat, 2023).

Starke KI oder künstliche *allgemeine* Intelligenz *(artificial general intelligence, AGI)* beschreibt ein System, dessen «Intelligenz» in allen Bereichen mit der menschlichen Intelligenz vergleichbar ist und demnach jede Aufgabe gleich gut oder besser als das menschliche Gehirn erfüllen kann (Goertzel, 2014). Diese starke KI existiert bislang lediglich in der Theorie (siehe aber z. B. Bubeck et al., 2023). Zudem ist es umstritten, ob es überhaupt möglich ist, eine solche starke KI jemals zu entwickeln (Fjelland, 2020).

Die Verwendung des Begriffes «Intelligenz» wird in diesem Zusammenhang teilweise kritisch gesehen, weil damit eine einseitig funktionalistische Sichtweise auf Intelligenz ausgedrückt werde. Der Funktionalismus beschreibt eine kausale Verknüpfung von Eingabe und Ausgabe. Ist ein System in der Lage, analog einem Menschen auf Basis einer bestimmten Eingabe eine bestimmte Ausgabe zu generieren, ist aus Sicht des Funktionalismus der Begriff «Intelligenz» auch bei diesem System angebracht, da sich Aus- und Eingabe nicht von derjenigen eines Menschen unterscheiden. Kritisiert wird daran, dass damit Aspekte wie die menschliche Vernunft, ein inhaltliches Verständnis und ein phänomenales Bewusstsein ausgeklammert werden (für eine vertiefte Diskussion, siehe Deutscher Ethikrat, 2023).

Maschinelles Lernen

Maschinelles Lernen (machine learning) ist ein Teilgebiet der KI. Beim maschinellen Lernen «übt» ein System anhand von Trainingsdaten, Muster zu erkennen und entwickelt darauf basierend Regeln *(Algorithmen).* Anhand dieser Regeln führt das System die Aufgabe dann selbständig aus (Banh & Strobel, 2023). Dieses Vorgehen bringt den Vorteil mit sich, dass dem Sys-

tem nicht mehr durch eine manuelle Programmierung bestimmte Regeln beigebracht werden müssen, wie es bei früheren Ansätzen notwendig war (Jordan & Mitchell, 2015). Neben dieser Effizienzsteigerung hat maschinelles Lernen auch den Vorteil, dass Regeln aufgebaut werden können, welche das System in den Trainingsdaten «entdeckt», welche den Programmierenden jedoch unbekannt geblieben wären. So kann auch die Erledigung der Aufgabe verbessert werden, weil nicht von Menschen vorgegebene, möglicherweise unvollständige oder fehlerhafte Regeln verwendet werden, sondern Regeln, welche auf den tatsächlichen Mustern in den Trainingsdaten basieren (Maggiori, 2023; Rebala et al., 2019).

Beim maschinellen Lernen werden verschiedene Formen unterschieden, wozu insbesondere das *überwachte Lernen (supervised learning),* das *unüberwachte Lernen (unsupervised learning)* und das *verstärkende Lernen (reinforcement learning)* gehören (Russell & Norvig, 2021).

Überwachtes Lernen *(supervised learning):*

Beim überwachten Lernen sind die Trainingsdaten mit einem *Label* versehen. Die korrekte Zuordnung von Eingabe *(Input)* und Ausgabe *(Output)* ist durch dieses Label festgelegt. Das System erkennt die zugrundeliegenden Muster und «übt» die korrekte Zuordnung. Dem System wird z. B. ein Bild einer Katze vorgelegt (Eingabe), welches mit dem Label «Katze» versehen wurde (Ausgabe). Das System «lernt» so mittels Wiederholung – d. h. der Eingabe von verschiedenen Bildern von Katzen –, wie eine Katze aussieht. Das Ziel ist es, dass das System auch Bilder von Katzen korrekt zuordnen kann, welche nicht Teil der Trainingsdaten waren.

Unüberwachtes Lernen *(unsupervised learning):*

Beim unüberwachten Lernen sind die Trainingsdaten nicht mit einem Label versehen. Das System muss selbst nach Mustern suchen und Regeln entwickeln, was meist mittels der Bildung von Kategorien geschieht *(clustering).* Durch wiederholtes Vorlegen verschiedener Bilder von Hunden und Katzen «lernt» das System, wie sich diese beiden Kategorien unterscheiden und auch unbekannte Bilder korrekt zuzuordnen.

Verstärkendes Lernen *(reinforcement learning):*

Beim verstärkenden Lernen soll sich das System mittels wiederholter Trainingsrunden immer mehr einem Ziel annähern. Das System erhält nach je-

der Trainingsrunde eine Rückmeldung in Form einer Verstärkung *(reinforcement)* oder einer Bestrafung *(punishment),* an welcher es das Verhalten in der darauf folgenden Trainingsrunde ausrichten kann (Russell & Norvig, 2021). Ist z. B. das Ziel, Bilder von Katzen und Hunden korrekt zu kategorisieren, wird dem System nach jeder Trainingsrunde rückgemeldet, ob es sich diesem Ziel angenähert hat – d. h. mehr korrekte Zuordnungen gemacht hat – oder nicht. Durch wiederholte Rückmeldungen kann das System die beiden Kategorien immer besser unterscheiden und mit der Zeit auch Bilder korrekt zuordnen, welche nicht Teil der Trainingsdaten waren.

Generative KI

Als *generative KI* werden KI-Systeme bezeichnet, welche komplexe neue Inhalte wie z. B. Texte, Bilder oder Videos ausgeben bzw. *generieren* können (Banh & Strobel, 2023). Generative KI basiert auf einem spezifischen Bereich des maschinellen Lernens, dem sog. *deep learning* (für einen Überblick, siehe z. B. LeCun et al., 2015; Sarker, 2021). Beim *deep learning* werden Informationen mittels einer Reihe von sog. *Schichten (layers)* verarbeitet, welche hierarchisch aufgebaut sind und unterschiedliche Aufgaben haben. Neben einer *Eingabeschicht (input layer)* und einer *Ausgabeschicht (output layer)* gibt es eine variable Anzahl an *Zwischenschichten (hidden layers),* welche eine komplexe Verarbeitung von Informationen ermöglichen. Im Zusammenhang mit diesen Schichten wird auch von *künstlichen neuronalen Netzwerken (artificial neural networks)* gesprochen, welche die Basis von *deep learning* bilden (Sarker, 2021).[3]

Deep learning kann sowohl als überwachtes Lernen als auch unüberwachtes Lernen stattfinden (LeCun et al., 2015). Weil dabei verschiedene Lernalgorithmen durch das System selbständig miteinander kombiniert werden, können im Vergleich zu anderen Formen von maschinellem Lernen auch deutlich komplexere Muster in grösseren Datenmengen identifiziert werden. *Deep learning* ermöglicht es damit, insbesondere auch unstrukturierte Inhalte wie z. B. Texte, Bilder oder Videos effizient zu verarbeiten (LeCun et al., 2015).

3 Die Verwendung des Begriffs *künstliche* neuronale Netzwerke lässt unschwer erkennen, dass damit eine Parallele zu den in menschlichen Gehirnen vorhandenen neuronalen Netzwerken gezogen werden soll.

Sprachmodelle

Zur generativen KI zählen u. a. die sog. *Sprachmodelle* (*large language models, LLMs*). Diese KI-Systeme sind darauf spezialisiert, Sprache zu verarbeiten und Texte zu generieren. Sprachmodelle werden auf Basis von unzähligen Textdaten trainiert, bestehend aus digital verfügbaren Texten. Dazu gehören z. B. frei verfügbare Webseiten, Online-Enzyklopädien (z. B. Wikipedia), digital verfügbare Bücher und wissenschaftlichen Publikationen (Liu et al., 2024).

Die Grundlage für Sprachmodelle besteht in der *Tokenisierung (tokenization).* Texte werden dabei in Worte beziehungsweise Wortteile unterteilt *(Tokens).* Für jedes dieser Tokens wird während des Trainingsprozesses «gelernt», wie häufig es mit bestimmten anderen Tokens verbunden ist bzw. auf bestimmte andere Tokens folgt (Wolfram, 2023). Die ermittelten Wahrscheinlichkeiten, mit denen Tokens aufeinander folgen, werden dann für die Generierung von neuem Text verwendet. Soll z. B. der Satz «das Gras ist …» von einem Sprachmodell ergänzt werden, wird dieses mit hoher Wahrscheinlichkeit das Wort «grün» ausgeben.

> "Contrary to how it may seem when we observe its output, an LM *[Language Model]* is a system for haphazardly stitching together sequences of linguistic forms it has observed in its vast training data, according to probabilistic information about how they combine, but without any reference to meaning: a stochastic parrot."
>
> (Im Gegensatz zu dem, was wir bei der Betrachtung seiner Ausgabe vermuten, ist ein Sprachmodell ein System, das willkürlich Sequenzen sprachlicher Formen zusammenfügt, die es in seinen riesigen Trainingsdaten beobachtet hat, und zwar auf der Grundlage probabilistischer Informationen darüber, wie sie kombiniert werden, aber ohne jeglichen Sinnbezug: ein stochastischer Papagei.)
>
> —Bender et al. (2021, pp. 616–617)

Das Sprachmodell «weiss» aber nicht, dass Gras in der Regel grün ist. Es hat lediglich «gelernt», dass im Zusammenhang mit «Gras» sehr häufig «grün» in Texten vorkommt (Waldo & Boussard, 2024). Damit sei auch schon eine erste wichtige Limitation von Sprachmodellen genannt: Sie wissen und verstehen nichts (Buck, 2025; Oertner, 2024). Sie sind aber sehr gut darin, verständliche und logisch klingende Sätze basierend auf stochastischen Zusammenhängen (d. h. Wahrscheinlichkeiten) von einzelnen Wortteilen zu generieren – sie sind «stochastische Papageien» (Bender et al., 2021).

Die Variabilität der von Sprachmodellen generierten Texte kommt aufgrund eines Parameters im Sprachmodell zustande, der sog. *temperature.* Die *temperature* beeinflusst die Wahrscheinlichkeiten für die Wahl der Tokens und weist üblicherweise Werte zwischen 0 und 2 auf. Bei einer *temperature* von 0 wird immer das Token mit der höchsten Wahrscheinlichkeit ausgewählt. Dies führt jedoch zu repetitiven Texten. Je höher die *temperature,* desto kreativer die Texte, weil auch Tokens mit geringerer Wahrscheinlichkeit ausgewählt werden. Ist die *temperature* hingegen zu hoch, generiert das Sprachmodell bei längeren Texten nur noch sinnfreie Wortreihen (Oertner, 2024; Wolfram, 2023).

KI-Tools

> «ChatGPT ist, bildlich gesprochen, die Karosserie, die verschiedenen Sprachmodelle aber der Motor des Autos.»
>
> —Buck (2025, p. 27)

Sprachmodelle bilden die Grundlage von Chatbots wie ChatGPT von OpenAI, Copilot von Microsoft, Claude von Anthropic oder DeepSeek vom gleichnamigen Anbieter. Chatbots ermöglichen es, anhand von Eingaben *(Prompts)* mit dem KI-System zu kommunizieren bzw. Ausgaben generieren zu lassen. Ein Prompt ist eine spezifische sprachliche Instruktion, welche an den Chatbot gerichtet ist und diesen zur Generierung einer bestimmten Ausgabe veranlassen soll (vgl. Prompting).

Chatbots können zusammen mit ähnlichen, anwendungsorientierten Tools ohne Chatfunktion (z. B. DeepL) unter dem Sammelbegriff «KI-Tools» subsumiert werden. Wenn im Folgenden von der Nutzung von KI zum Schreiben die Rede ist, sind insbesondere diese KI-Tools sowie deren zugrundeliegenden Sprachmodelle gemeint.

Anthropomorphisierung

> «Die Anthropomorphisierung digitaler Technik ist in der Umgangssprache weit fortgeschritten. Sie zeigt sich darin, dass KI und Robotern Fähigkeiten wie Denken, Lernen, Entscheiden oder Emotionalität zugeschrieben werden, wodurch sie scheinbar in die Gemeinschaft der denkenden, lernenden, entscheidenden und fühlenden Menschen aufgenommen werden.»
>
> —Deutscher Ethikrat (2023, p. 170)

Aufmerksamen Lesenden wird aufgefallen sein, dass in den vorherigen Abschnitten gewisse Verben in Anführungszeichen gesetzt wurden, z. B. das Sprachmodell «weiss» oder hat «gelernt». Diese Zuschreibung von menschlichen Eigenschaften auf Nichtmenschliches wird *Anthropomorphisierung* oder *Anthropomorphismus* genannt (Airenti, 2015; Epley et al., 2007). Eine damit verwandte Begrifflichkeit ist der *Animismus.* Der Animismus ist weiter gefasst als der Anthropomorphismus und meint allgemein die Zuschreibung von Leben auf nichtlebende Objekte (Epley et al., 2007). Im religiösen Kontext wird damit insbesondere die Vorstellung verstanden, dass alle Objekte – also nicht nur Lebewesen – eine Seele besitzen (Schlatter, 2005).

> "Anthropomorphic language is so prevalent in the discipline that it seems inescapable. Perhaps part of the reason is because anthropomorphism is built, analytically, into the very concept of AI. The name of the field alone—artificial intelligence—conjures expectations by attributing a human characteristic—intelligence—to a non-living, non-human entity, which thereby exposes underlying assumptions about the capabilities of AI systems."
>
> (Die anthropomorphe Sprache ist in diesem Fachgebiet so weit verbreitet, dass sie unausweichlich erscheint. Ein Grund dafür könnte sein, dass Anthropomorphismus analytisch gesehen zum Kernkonzept der KI gehört. Allein schon der Name des Fachgebiets – künstliche Intelligenz – weckt Erwartungen, indem er einem nicht lebenden, nicht menschlichen Wesen eine menschliche Eigenschaft – Intelligenz – zuschreibt, wodurch grundlegende Annahmen über die Fähigkeiten von KI-Systemen offenbart werden.)
>
> —Placani (2024, p. 692)

Die Tendenz zur Anthropomorphisierung zeigt sich im Zusammenhang mit KI nur schon an der Begrifflichkeit «Künstliche Intelligenz». Diese metaphorische Verwendung zeugt von einer ungerechtfertigten Vereinfachung des Konzepts der Intelligenz und fördert das Missverständnis, es handle sich bei menschlicher Intelligenz und KI um vergleichbare oder zumindest ähnliche Konzepte (Deutscher Ethikrat, 2023).

> «Softwaresysteme, auch solche, die der KI zugerechnet werden, verfügen weder über theoretische noch über praktische Vernunft. Sie können keine Verantwortung für ihr Handeln übernehmen, sie sind kein personales Gegenüber, auch dann nicht, wenn sie Anteilnahme, Kooperationsbereitschaft oder Einsichtsfähigkeit simulieren.»
>
> —Deutscher Ethikrat (2023, p. 334)

KI-Tools, welche mittels menschlicher Sprache scheinbar – und in eloquenter Art und Weise – auf ein Gegenüber eingehen können, fördern die Tendenz zur Anthropomorphisierung besonders, weil dieses Verhalten bei Menschen zu einem Erleben von Verbundenheit und Empathie führt (Airenti, 2015; Placani, 2024).[4] Je mehr sich ein KI-Tool somit wie ein menschliches Gegenüber «verhält», desto eher wird es auch fälschlicherweise für ein menschliches Gegenüber gehalten (Peter et al., 2025).

Allerdings ist die Verwendung von anthropomorphisierenden Zuschreibungen bei KI-Tools nicht nur faktisch falsch, sondern sogar gefährlich. Dies kann zu einem falschen Verständnis der Funktionsweise von KI-Tools beitragen und zu deren Überschätzung führen (Gredel et al., 2024; Hicks et al., 2024; Meier-Vieracker, 2024). Eine solche Überschätzung von KI-Tools zeigt sich nicht nur, wenn den Ausgaben von KI-Tools unkritisch Glauben geschenkt wird, sondern auch, wenn Nutzende der Illusion erliegen, KI-Tools würden über zwischenmenschliche Kompetenzen verfügen oder könnten Gefühle entwickeln – z. B. wenn KI-Tools für empathische und verständnisvolle Therapeut:innen gehalten werden (Poorghadiri, 2025) oder sogar romantische Gefühle für ein KI-Tool aufkommen (Ebner & Szczuka, 2025). Dementsprechend wichtig ist das Bewusstsein darüber, dass Menschen diese Neigung zur Vermenschlichung haben und diese durch sprachbasierte KI-Tools nochmals deutlich verstärkt wird.

---

4 Dies liegt u. a. auch daran, dass diese KI-Tools auf Basis von durch Menschen erschaffenen Texten funktionieren.

Ethische und rechtliche Rahmenbedingungen

Autorenschaft und Verantwortung

> "ChatGPT is fun, but not an author."
> (ChatGPT macht Spass, ist aber keine Autor:in.)
>
> —Thorp (2023, p. 313)

Mit der Verbreitung von ChatGPT rückte eine Frage in den Fokus, die sich mit den Ausgaben solcher KI-Tools befasst: Können KI-Tools als Autor:innen ihrer Ausgaben betrachtet werden? Wer ist Urherber:in von den mit KI-Tools erstellten Texten – menschliche Autor:innen, KI-Tools oder beide? Insbesondere im wissenschaftlichen Kontext, wo Urheberschaft und der Angabe von Quellen im Zusammenhang mit der wissenschaftlichen Integrität viel Bedeutung beigemessen wird, wurden solche Fragen schnell aufgegriffen. Einige frühe Artikel, die mithilfe von ChatGPT erstellt wurden, wiesen das Tool sogar explizit als Autor:in aus (z. B. Benichou & ChatGPT, 2023; Curtis & ChatGPT, 2023). Als Reaktion auf diese Praxis wurden in vielen wissenschaftlichen Fachzeitschriften Richtlinien zum Umgang mit KI-Tools veröffentlicht, nach welchen nur Menschen als Autor:innen aufgeführt werden dürfen.

> "No LLM tool will be accepted as a credited author on a research paper. That is because any attribution of authorship carries with it accountability for the work, and AI tools cannot take such responsibility."
> (Kein LLM-Tool wird als anerkannte Autor:in einer Forschungsarbeit akzeptiert. Der Grund dafür ist, dass jede Zuschreibung von Autorenschaft eine Verantwortung für die Arbeit mit sich bringt, und KI-Tools können diese Verantwortung nicht übernehmen.)
>
> —Editorial in der wissenschaftlichen Fachzeitschrift *Nature* ("Tools Such as ChatGPT Threaten Transparent Science; Here Are Our Ground Rules for Their Use," 2023, p. 612).

Diese Sichtweise, dass KI-Tools Instrumente sind und keine Autoren- oder Urheberschaft beanspruchen können, deckt sich auch mit Rechtsgutachten im Zusammenhang mit der Nutzung von KI (z. B. Salden & Leschke, 2023; United States Copyright Office, 2025a).

Damit einhergehend wird auch direkt die Frage nach der Verantwortung für die mittels KI-Tools erstellten Texte beantwortet (vgl. Deutscher Ethik-

rat, 2023). Werden KI-Tools als Instrumente bzw. Werkzeuge bei der Erstellung von Texten genutzt, liegt die Verantwortung für die Integrität und Korrektheit der Ergebnisse vollumfänglich bei den Autor:innen. Ähnlich wie z. B. eine Handwerkerin nicht ihrem Hammer die Schuld für eine fehlerhafte Montage geben wird, können auch Autor:innen nicht einem KI-Tool die Schuld für einen fehlerhaften Text zuschieben. Mit der Rechtfertigung, ein Fehler sei auf ein KI-Tool zurückzuführen, überführen sich die Autor:innen somit ggf. selbst einer (wissenschaftlichen) Unredlichkeit.

Deklaration

> "You may not use our Services for any illegal, harmful, or abusive activity. For example, you are prohibited from … representing that Output *[sic]* was human-generated when it was not."
>
> (Sie dürfen unsere Dienste nicht für illegale, schädliche oder missbräuchliche Aktivitäten nutzen. So ist es Ihnen beispielsweise untersagt zu behaupten, dass die Ausgabe von Menschen erstellt wurde, wenn dies nicht der Fall war.)
>
> —OpenAI Nutzungsbestimmungen (OpenAI, 2025b, Chapter Using our Services)

Ist die Verwendung eines KI-Tools im Rahmen einer Schreibaufgabe erlaubt[5], so muss dies deklariert werden. Dies gebietet zum einen – zumindest im wissenschaftlichen Kontext – bereits die wissenschaftliche Integrität. Zum anderen setzen KI-Tools in den Nutzungsbestimmungen *(terms of use)* vielfach voraus, dass die Leserschaft darüber informiert werden muss, wenn ein Text durch das KI-Tool generiert wurde.

Die Verwendung von KI-Tools erfordert, dass im Text gekennzeichnet ist, welche Stellen (zumindest teilweise) durch ein KI-Tool generiert wurden – inkl. Angaben zur Art und Version des KI-Tools und dessen Nutzung (z. B. Prompts). Da die von KI-Tools generierten Inhalte für die Leserschaft nicht auffindbar bzw. reproduzierbar sind, ist eine Angabe zum generierten Text im Literaturverzeichnis jedoch nicht sinnvoll. Das KI-Tool selbst kann allerdings zumindest als Software auch im Literaturverzeichnis zitiert werden – inkl. Angaben zur Art und Version der Software. Hier als Beispiel für ChatGPT dargestellt:

5 Im Rahmen von z. B. Prüfungen und Qualifikationsarbeiten an Hochschulen gelten häufig Bestimmungen, welche über diese rechtlichen und wissenschaftsethischen Aspekte hinausgehen.

Open AI. (2025). ChatGPT (Version July 25) [Generative pre-trained transformer]. https://chatgpt.com/

Werden KI-Tools hingegen ausschliesslich zur Verbesserung der Grammatik – inkl. Syntax und Rechtschreibung – verwendet, muss dies in der Regel nicht transparent gemacht werden, da entsprechende Programme in klassischen Textverarbeitungsprogrammen ebenfalls bereits integriert sind und standardmässig genutzt werden. Ähnlich wie es heute normal ist, dass z. B. Schüler:innen Aufsätze an PCs – mit Programmen, welche die Grammatik überprüfen – schreiben und nicht mehr von Hand, wird es zukünftig normal sein, dass Texte von KI auf Grammatik geprüft werden.

Auch Hilfestellungen durch KI-Tools, welche nicht die eigentliche «Arbeit am Text» betreffen (z. B. Informationssuche, Ideensammlung), müssen nicht zwingend transparent gemacht werden, da z. B. menschliche Unterstützung gleicher Art ebenfalls nicht spezifisch ausgewiesen wird. Allerdings empfiehlt es sich insbesondere für wissenschaftliche Publikationen und Qualifikationsarbeiten an Hochschulen, die konkreten Vorgaben von wissenschaftlichen Fachzeitschriften bzw. Hochschulen zu konsultieren.

Eine Erkennung von mittels KI-Tools generiertem Text ist nach aktuellem Stand nicht zweifelsfrei möglich (für eine Übersicht, siehe Wu et al., 2024). Verschiedene Studien zeigen, dass die Detektionsrate von KI-generiertem Text sowohl bei Fachpersonen (Casal & Kessler, 2023; Fleckenstein et al., 2024; Gao et al., 2023; Waltzer et al., 2024) als auch bei vielen spezifisch dafür entwickelten Detektionstools (Walters, 2023; Weber-Wulff et al., 2023) ausbaufähig ist. Doch selbst wenn eine zuverlässige Erkennung von durch KI-generierten Texten in ihrer Grundform möglich wäre, sind die Möglichkeiten begrenzt. Wird der Text z. B. nach der Generierung durch ein KI-Tool umformuliert oder werden gewisse Wörter ersetzt, kann die korrekte Erkennung als KI-generiert dennoch erschwert oder verunmöglicht werden (Anderson et al., 2023; Sadasivan et al., 2025; Wu et al., 2024).

Eine weitere Möglichkeit zur Erkennung besteht in der Markierung von generiertem Text mittels Wasserzeichen *(watermarking)*. Solche Wasserzeichen lassen sich entweder im generierten Text selbst[6] (Dathathri et al., 2024) oder im Quellcode des Textes (Koch, 2025) einbetten. Jedoch können auch Wasserzeichen umgangen werden, indem der generierte Text umfor-

6 D.h. der Text weist ein bestimmtes algorithmisches Muster auf, welches von einem Detektionstool erkannt wird.

muliert (siehe oben) oder der Quellcode des Textes entfernt wird.[7] Zudem besteht bei solchen Wasserzeichen umgekehrt das Problem, dass selbst geschriebener Text allenfalls als KI-generiert markiert wird, wenn der Text für leichtfügige Überarbeitungen in ein KI-Tool hochgeladen und eine überarbeitete – mit einem Wasserzeichen versehene – Ausgabe aus dem KI-Tool herauskopiert wird (Koch, 2025).

Insgesamt ist somit die Erkennungsrate zu tief bzw. die Falsch-Positiv-Rate zu hoch, sodass solche Detektionstools nicht zur zuverlässigen Erkennung von KI-generiertem Text verwendet werden können.

Urheberrecht

Über Autorenschaft, Verantwortung und Deklaration hinaus gibt es weitere Aspekte, welche im Rahmen des rechtlichen und ethisch verantwortungsvollen Umgangs mit KI-Tools berücksichtigt werden müssen (Zhou et al., 2024). Ein erster Punkt betrifft die Wahrung des *Urheberrechts* (Fang & Perkins, 2024; Karamolegkou et al., 2023). Bereits existierende Texte sind oftmals urheberrechtlich geschützt und dürfen daher nicht ohne die Erlaubnis der Urheber:innen – oder anderer die Rechte innehabender Personen oder Körperschaften – verwendet werden. Verschiedene KI-Unternehmen wurden dafür kritisiert, dass sie urheberrechtlich geschützte Texte ohne eine entsprechende Erlaubnis in die Trainingsdaten ihrer Sprachmodelle mit aufnahmen. In diesem Zusammenhang kam es auch schon zu mehreren Schadenersatzklagen (Bendel, 2024; Buck, 2025).[8]

Aktuell stellen sich z. B. die US-Anbieter von KI-Tools jedoch auf den Standpunkt, dass sie aufgrund der in den USA geltenden *"fair use"* Doktrin[9] auch urheberrechtlich geschützte Texte als Trainingsdaten verwenden dür-

7 Z. B. mittels Zwischenkopieren in einen Texteditor. Es gibt auch Tools, mit welchen der Quellcode sichtbar gemacht werden kann, z. B. von Soscisurvey oder Originality.ai.

8 Bei den Klagen geht es jeweils um die Trainingsdaten und nicht um die Ausgaben von KI-Tools. Die Texte in den Trainingsdaten werden aufgrund der Funktionsweise der KI-Tools nicht eins zu eins ausgegeben. Die Wahrscheinlichkeit, dass z. B. ein Buchkapitel exakt so ausgegeben wird, wie es in den Trainingsdaten vorhanden ist, geht gegen null.

9 Die *"fair use"* Doktrin besagt, dass urheberrechtlich geschützte Werke (z. B. Texte) in gewissen Fällen auch ohne Einwilligung der Urheber:innen genutzt werden dürfen, allenfalls sogar von kommerziellen Unternehmen. Dabei wird im Einzelfall jeweils eine Abwägung der Interessen der betroffenen Parteien getroffen. Als mögliche Gründe für *"fair use"* im Zusammenhang mit KI-Tools gelten z. B. eine transformative Verwendung (d.h. Erschaffung von etwas Neuem) und der sich ergebende gesellschaftliche Nutzen (United States Copyright Office, 2025b).

fen. Aus rechtlicher Sicht ist es aber fraglich, ob dieser *"fair use"* immer gegeben ist oder ob für die Verwendung als Trainingsdaten z. B. eine explizite Einwilligung von Autor:innen von urheberrechtliche geschützten Texten eingeholt werden muss (United States Copyright Office, 2025b). In der EU haben Autor:innen durch den 2025 in Kraft gesetzten *Artificial Intelligence Act* gemäss gängiger Meinung lediglich das Recht, die Verwendung ihrer Texte für Trainingsdaten zu untersagen *(opt-out)* – ohne dass aktuell geklärt ist, in welcher Form dies geschehen soll. Diese Regelung wurde bereits verschiedentlich kritisiert und das letzte Wort scheint noch nicht gesprochen bzw. geschrieben worden zu sein (Dornis & Stober, 2024; Rankin, 2025; United States Copyright Office, 2025b).

Darüber hinaus ist das Urheberrecht auch auf individueller Ebene relevant. Wenn eine Person bei der Nutzung eines KI-Tools einen urheberrechtlich geschützten Text ohne Erlaubnis in das Tool hochlädt, begeht sie allenfalls ebenfalls eine Urheberrechtsverletzung (Buck, 2025). Um dies zu verhindern, lässt sich in KI-Tools jedoch üblicherweise einstellen, ob die mit dem KI-Tool geteilten Inhalte für Trainingszwecke verwendet werden dürfen. Andernfalls müsste immer geprüft werden, ob ein Text überhaupt in ein KI-Tool hochgeladen werden darf – in der Praxis dürfte dies kaum je geschehen.

Datenschutz

> "A major risk is that user-provided information will be incorporated into the training data of proprietary LLMs without user consent, and that this will result in the leaking of personally identifiable information and/or other data that users would want to remain private."
>
> (Ein grosses Risiko besteht darin, dass von Nutzenden zur Verfügung gestellte Informationen ohne deren Zustimmung in die Trainingsdaten unternehmenseigener Sprachmodelle (LLMs) aufgenommen werden und dass dies zur Offenlegung persönlich identifizierbarer Informationen und/oder anderer Daten führt, von denen die Nutzenden möchten, dass diese privat bleiben.)
>
> —Fang und Perkins (2024, p. 136)

Ein weiterer Punkt ist die Wahrung des *Datenschutzes*. Werden eigene oder fremde Daten, Bilder oder Dokumente in ein KI-Tool hochgeladen – oder sind diese bereits in den Trainingsdaten vorhanden –, können diese durch den Anbieter für Zwecke verwendet werden, welche von den betroffenen

Personen nicht angedacht waren (Buck, 2025; Fang & Perkins, 2024; Kshetri, 2023). Auch die eingegebenen Prompts können – je nach Einstellungen im KI-Tool – vom Anbieter als Informationsquelle genutzt werden (Bendel, 2024). Darüber hinaus besteht die Gefahr, dass Daten durch Datenlecks in falsche Hände gelangen (B. Yan et al., 2025). Immer wieder werden z. B. Fälle publik, bei denen unternehmensinterne Informationen durch die Nutzung von KI-Tools geleakt (Ray, 2023) oder vermeintlich private Konversationen mit KI-Tools veröffentlicht wurden (Rezaee, 2025).

Um solche negativen Auswirkungen zu verhindern, sollte mit eigenen und insbesondere mit fremden Daten umsichtig umgegangen werden.V. a. personenbezogene Daten (z. B. Namen, Kontaktdaten) und unternehmensinterne Informationen sollten nicht in KI-Tools eingegeben werden, wenn nicht sichergestellt ist, dass diese nicht für andere als die beabsichtigten Zwecke verwendet werden (Buck, 2025).[10]

Schädliche Verwendung

> "Individuals without any programming skills and writing skills can engage in cybercrimes. The only thing they need is the skill to write good prompts."
>
> (Personen ohne Programmier- und Schreibfähigkeiten können Cyberverbrechen begehen. Das Einzige, was sie brauchen, ist die Fähigkeit, gute Prompts schreiben zu können.)
>
> —Kshetri (2023, p. 12)

Es besteht die Gefahr, dass KI-Tools für kriminelle Zwecke genutzt werden (*Cyberkriminalität;* Kshetri, 2023). Mögliche Verwendungsbereiche sind z. B. *Social Engineering,* die *Programmierung von Schadsoftware* und *Hacking,* welche alle durch KI-Tools deutlich erleichtert werden (Brundage et al., 2018; Kshetri, 2023; Oertner, 2024). Social Engineering meint das Vorspielen einer falschen Identität, um sich z. B. Zugang zu sensiblen Daten oder geschützten Netzwerken zu verschaffen. Eine weit verbreitete Form des Social Engineering ist *Phishing,* bei welchem z. B. mithilfe von gefälschten E-Mails versucht wird, an sensible Daten oder Passwörter der Adressat:innen zu gelangen. Social Engineering ist durch die Generierung von KI-In-

10 Ein Blick in die Nutzungsbedingungen *(terms of use)* und Datenschutzbestimmungen *(privacy policy)* ist empfehlenswert (Kshetri, 2023).

halten in Text, Ton, Bild und/oder Video viel einfacher möglich und durch eine höhere wahrgenommene Authentizität zudem erfolgsversprechender als mit bisherigen Mitteln (Dwivedi et al., 2023; Kshetri, 2023). Die Programmierung von Schadsoftware und Hacking wird erleichtert, da nutzbarer Programmiercode direkt durch KI-Tools generiert werden kann (Kshetri, 2023).

Neben diesen Aspekten, welche die digitale Sicherheit betreffen, gibt es auch Aspekte, welche die politische Sicherheit betreffen. Dazu gehört u. a. die bewusste Verbreitung von falschen Informationen mit Täuschungsabsicht *(Desinformation).* Z. B. können fabrizierte Texte, Bilder oder Videos einfacher erstellt und veröffentlicht werden. Auch eine gezielte Desinformation von Einzelpersonen durch personalisierte Nachrichten ist durch KI-Tools viel einfacher und in grösserem Umfang möglich (Brundage et al., 2018; Deutscher Ethikrat, 2023; Fang & Perkins, 2024).

Zwar sind in vielen KI-Tools Restriktionen in Form von Filtern einprogrammiert, welche eine kriminelle oder unethische Verwendung verhindern sollen. Wird ein Prompt vom KI-Tool als Verstoss gegen die Nutzungsrichtlinien klassifiziert, weist das Tool darauf hin und gibt in der Folge keinen Text aus. Jedoch sind solche Schutzmechanismen nicht perfekt und werden regelmässig von Cyberkriminellen umgangen (Kshetri, 2023; Oertner, 2024). Darüber hinaus drängt sich bei diesen programmierten Restriktionen die Frage auf, ob sie immer gerechtfertigt sind. Geht es z. B. nicht um rechtliche Einschränkungen, sondern lediglich um durch den Anbieter des Tools definierte moralische Regeln, könnte das KI-Tool – im Sinne einer Zensur – Ausgaben «verweigern», auch wenn dies aus gesellschaftlicher oder individueller Sicht nicht gerechtfertigt wäre (Bendel, 2024). So gibt z. B. das chinesische KI-Tool DeepSeek keine Antwort auf bestimmte menschenrechtsbezogene Ereignisse im Kontext der chinesischen Politik aus (Lu, 2025).

7.2 Chancen und Risiken beim Schreiben mit KI

Inhaltsübersicht

Einleitung

Die Nutzung von KI-Tools beim Schreiben birgt sowohl erhebliche Chancen als auch Risiken. In diesem Kapitel werden verschiedene Aspekte im Zusammenhang mit dem Schreiben beleuchtet, bei welchen sich KI-Tools positiv und/oder negativ auswirken können (vgl. Albrecht, 2023; Buck, 2025; Friedrich et al., 2024; Gredel et al., 2024; Morrison, 2023; Oertner, 2024; Weidinger et al., 2022). Der Reihe nach wird dabei eingegangen auf die Aspekte (a) Effizienz, (b) Textqualität, (c) Transparenz, (d) Kompetenzerwerb sowie (e) Handlungsfähigkeit und Motivation beim Schreiben.

Effizienz

Chancen

> "Generative AI tools like Copilot and ChatGPT can boost writing productivity by assisting with tasks such as content generation, idea creation, and stylistic editing, helping both expert and novice writers."
> (Generative KI-Tools wie Copilot und ChatGPT können die Produktivität beim Schreiben steigern, indem sie bei Aufgaben wie der Generierung von Inhalten, der Ideenfindung und der stilistischen Überarbeitung unterstützen und dadurch sowohl erfahrenen als auch unerfahrenen Schreibenden helfen.)
>
> —H.-P. Lee et al. (2025, Section 2.3)

Die augenscheinlich grösste Chance der Nutzung von KI-Tools beim Schreiben besteht in einer Erhöhung der Effizienz. Schreibaufgaben können ganz oder teilweise an KI-Tools ausgelagert werden. Als Konsequenz benötigen diese Aufgaben weniger zeitliche und kognitive Ressourcen (u. a. geringerer *cognitive load;* Stadler et al., 2024). Dabei bietet sich insbesondere die Auslagerung von Routineaufgaben an, welche weniger menschliche «Überwachung» benötigen (Dobrin, 2023; Rafner et al., 2021). Dazu gehören auch gänzlich durch KI-Tools generierte kurze Texte, welche für gewisse Nutzungsbereiche bereits verbreitet sind, wie z. B. Bildbeschreibungen oder kürzere Nachrichtentexte von Nachrichtenagenturen (Becker, 2024). Im wissenschaftlichen Kontext zeigte sich z. B., dass es bereits möglich ist, *Abstracts* (d. h. Zusammenfassungen wissenschaftlicher Artikel) durch KI-Tools generieren zu lassen, welche sich qualitativ nur geringfügig von durch Menschen geschriebenen Abstracts unterscheiden (Gao et al., 2023).

Resultate verschiedener Studien untermauern den positiven Effekt von KI-Tools auf die Effizienz (z. B. Dell'Acqua et al., 2023; Noy & Zhang, 2023). Teilweise zeigt sich darüber hinaus, dass der positive Effekt auf die Effizienz bei geringeren Kompetenzen grösser ist als bei ausgereiften Kompetenzen (Noy & Zhang, 2023). Schreibende mit geringer Schreibkompetenz könnten demnach von KI-Tools besonders profitieren, um ihre Effizienz beim Schreiben zu erhöhen.

Risiken

Eine effiziente Nutzung von KI-Tools ist jedoch nur dann möglich, wenn Nutzende im Umgang mit KI-Tools geübt sind und die Aufgabe zum Anwendungsbereich von KI-Tools passt. Ansonsten könnte es sein, dass die Aufgabe sogar weniger effizient bearbeitet wird als ohne KI-Tools (Becker et al., 2025) oder dass der effizienzsteigernde Effekt zumindest ausbleibt (Bašić et al., 2023). Wird im Vorhinein eine Steigerung der Effizienz erwartet bzw. eingeplant, besteht somit das Risiko, dass die Umsetzung des Schreibprojekts negativ beeinflusst und sich damit schlimmstenfalls auch die Textqualität verringert.

Zentral ist zudem, dass die Effizienzsteigerung beim Schreiben eines Textes nicht zu Lasten der Qualität geht. Ein Teil der frei gewordenen Ressourcen muss demnach für eine inhaltliche Überprüfung des mithilfe von KI-Tools erstellten Textes genutzt werden (Kacena et al., 2024). Dies bedingt jedoch eine grundlegende Lesekompetenz und Informationskompetenz bei den Schreibenden. Sind diese Kompetenzen nicht bzw. nicht genügend vorhanden, führt dies dazu, dass etwas «geschrieben» wird, was gar nicht verstanden wird.

Textqualität

Chancen

> "By offloading to the AI some of the mental work …, writers can free up cognitive resources for other tasks. Cognitive offloading therefore improves performance and reduces the likelihood of errors."
>
> (Indem ein Teil der Denkarbeit an die KI abgegeben wird, können Autor:innen kognitive Ressourcen für andere Aufgaben freisetzen. Das «kognitive Abladen» verbessert daher die Leistung und verringert die Wahrscheinlichkeit von Fehlern.)
>
> —Z. Lin (2024b, p. 426)

Mit der Erhöhung der Effizienz kann auch eine Erhöhung der Textqualität einhergehen. Dadurch werden *zeitliche und kognitive Ressourcen* für andere Aufgaben frei, wie z. B. für Rechercheaufgaben oder für die Verbesserung der inhaltlichen Tiefe eines Textes (Buck, 2025; Z. Lin, 2024b).

> "It is comparatively easy to make computers exhibit adult-level performance in solving problems on intelligence tests or playing checkers, and difficult or impossible to give them the skills of a one-year-old when it comes to perception and mobility."
>
> (Es ist vergleichsweise einfach, Computer dazu zu bringen, beim Lösen von Problemen in Intelligenztests oder beim Damespiel die Leistung eines Erwachsenen zu erbringen und schwierig oder unmöglich, ihnen die Kompetenzen eines Einjährigen in Bezug auf Wahrnehmung und Mobilität zu verleihen.)
>
> —Moravec (1988, p. 15)

Darüber hinaus können KI-Tools die Textqualität auch auf anderem Weg positiv beeinflussen. Diese positiven Effekte werden unter dem Begriff der *hybriden Intelligenz* (Dellermann et al., 2019; Rafner et al., 2021) zusammengefasst. Damit ist eine Kombination der Kompetenzen von KI-Tools (z. B. Informationsverarbeitung) und menschlichen Nutzenden (z. B. Wahrnehmung) gemeint, bei welcher sich die Stärken der beiden so ergänzen, dass ein optimales Resultat erzielt wird. Die Verwendung von KI-Tools führt so z. B. durch Generierung von neuen Ideen und Perspektiven zu höherer Kreativität und besserer Argumentation (Doshi & Hauser, 2024; B. C. Lee & Chung, 2024; Marji & Licato, 2025). Auch die Verständlichkeit von Texten kann durch KI-Tools verbessert werden (Buck, 2025; Patel et al., 2023).

Eine zusätzliche Qualitätssteigerung ist bei Prompt-basierten KI-Tools durch den *«Zwang» des Promptings* möglich. Um einen Prompt genau formulieren und so eine nutzbare Ausgabe zu erhalten, wird bei den Schreibenden eine tiefere Reflexion darüber angeregt, was sie eigentlich schreiben möchten. Die aus dieser Reflexion gewonnenen Einsichten können sich dann wiederum positiv auf die Textqualität auswirken (Buck, 2025).

Analog zur Effizienz zeigen auch zur Textqualität verschiedene Studien positive Effekte der Nutzung von KI-Tools (z. B. Doshi & Hauser, 2024; Nguyen et al., 2024; Noy & Zhang, 2023). Auch hier profitieren Schreibende mit geringerer Schreibkompetenz allenfalls mehr von der Unterstützung durch KI-Tools, um die Textqualität zu erhöhen (Noy & Zhang, 2023).

Risiken

In Bezug auf die Textqualität ergeben sich auch verschiedene Risiken. Einerseits kann durch eine suboptimale Nutzung von KI-Tools ein positiver

Effekt auf die Textqualität ausbleiben oder diese wird sogar verringert. Denkbar sind u. a. folgende Arten der suboptimalen Nutzung:

1. Unpassende Aufgabenwahl: Aufgaben, welche (gemäss Experteneinschätzung) nicht im Kompetenzbereich des KI-Tools liegen, werden weniger gut erledigt als ohne KI-Unterstützung (Dell'Acqua et al., 2023).
2. Geringe Schreibkompetenz: Je nach Aufgabe führt eine wenig ausgereifte Schreibkompetenz dazu, dass das Potential von KI-Tools nicht ausgeschöpft wird und sich daher keine Qualitätssteigerung einstellt (Bašić et al., 2023).
3. Geringe Kompetenz im Umgang mit KI-Tools *(AI literacy):* Auch bei geeigneten Aufgaben ist der Umgang mit dem KI-Tool entscheidend.
   - Wird ein KI-Tool lediglich als Informationsquelle genutzt, führt dies in der Regel zu einer geringeren Textqualität als wenn ein KI-Tool mittels einer wiederholt stattfindenden Prozessschlaufe in den eigenen Schreibprozess integriert wird (Nguyen et al., 2024).
   - Verfügen Nutzende nicht über die Kompetenz, die Ausgabe einschätzen bzw. kritisch zu hinterfragen, sind oftmals ebenfalls Qualitätseinbussen die Folge.

Andererseits können bei der Nutzung von KI-Tools verschiedene Probleme auftreten, welche einen negativen Effekt auf die Textqualität haben. Darunter fallen verschiedene Aspekte wie (a) inhaltliche Fehler, (b) Verzerrungen *(biases),* (c) fehlende Originalität und inhaltliche Tiefe, (d) Verwässerung des eigenen Schreibstils sowie (e) sprachliche Fehler und Inkonsistenzen. Da das Bewusstsein über diese Risiken und deren Ursachen wesentlich dazu beitragen kann, negative Auswirkungen auf die Qualität des eigenen Schreibens zu verhindern, werden diese fünf Aspekte nachfolgend genauer erläutert.

Inhaltliche Fehler:

> "Essentially, ChatGPT can largely be understood as the code version of the 'infinite monkeys theorem,' in which it is surmised that, given enough time, an infinite number of monkeys bashing away at an infinite number of typewriters will … reproduce, say, the works of Shakespeare. It's not magic; it's just grunt work made workable by the sheer scale of resources thrown at it, with the resources in the case of generative AI being of the computing rather than simian variety."

> (Im Wesentlichen kann ChatGPT als die Code-Version des «Theorems der endlos tippenden Affen» verstanden werden, bei welchem man davon ausgeht, dass bei genügend Zeit eine unendliche Anzahl von Affen, die auf eine unendliche Anzahl von Schreibmaschinen einhämmern, z. B. die Werke von Shakespeare reproduzieren können. Das ist keine Magie; das ist nur Routinearbeit, die durch die schiere Menge der eingesetzten Ressourcen machbar wird, wobei die Ressourcen im Fall der generativen KI eher von der Sorte Computer als von der Sorte Affen sind.)
>
> —Morrison (2023, p. 157)

KI-Tools verstehen nichts, sie sind lediglich sehr gut darin, plausibel klingende Texte zu generieren (Friedrich et al., 2024; Morrison, 2023). KI-Tools wurden auch nicht dafür programmiert, korrekte Aussagen zu generieren. Daher ist es nicht weiter erstaunlich, dass es unzählige Beispiele von Fehlern in Ausgaben von KI-Tools gibt (siehe z. B. Y. Sun et al., 2024). Weshalb es zu diesen Fehlern kommt, hat unterschiedliche Gründe, welche z. T. sogar den Anbietern von KI-Tools unbekannt sind (OpenAI, 2025a). Die Hauptgründe hängen jedoch alle mit der Funktionsweise dieser KI-Tools zusammen.

Möglich sind einerseits fehlerhafte Ausgaben, welche *aufgrund der Trainingsdaten* entstehen (Albrecht, 2023; Deutscher Ethikrat, 2023). Dazu gehören weit verbreitete Fehlannahmen. Ein Beispiel für eine solche weit verbreitete Fehlannahme wäre, dass Menschen nur 10 Prozent ihres Gehirns nutzen. Weit verbreitete Fehlannahmen kommen in den Trainingsdaten häufiger vor und werden somit mit höherer Wahrscheinlichkeit vom KI-Tool wiedergegeben. Daneben bestehen die Trainingsdaten nicht nur aus vertrauenswürdigen Quellen wie z. B. offiziellen Dokumenten oder wissenschaftlichen Publikationen, sondern auch z. B. aus Blogposts oder Beiträgen in Online-Foren (Waldo & Boussard, 2024). Ein KI-Tool kann jedoch die zugrundeliegenden Daten nicht auf deren Wahrheitsgehalt überprüfen oder gewichten – alle Informationen in den Trainingsdaten sind gleichwertig, unabhängig davon wie vertrauenswürdig die Quelle ist, aus welcher die Informationen stammen (Kacena et al., 2024; Oertner, 2024). Bei Fehlern aufgrund der Trainingsdaten ist daher das in diesem Zusammenhang häufig genannte Schlagwort *"garbage in, garbage out"* sehr treffend.

Entsprechend wichtig ist eine hohe Qualität (im Sinne von Akkuratheit und Informationsbreite) der Trainingsdaten, damit es zu möglichst wenigen fehlerhaften Ausgaben kommt. Durch grössere Datensätze lassen sich solche Probleme nicht lösen, weil auch bei grossen, gut angelernten Sprachmodellen teilweise Fehler wiedergegeben werden, wenn diese häufig in den

Trainingsdaten vorkommen (S. Lin et al., 2022). Grössere Sprachmodelle bergen umgekehrt sogar das Risiko, dass mehr fehlerhafte Ausgaben erfolgen: Dadurch, dass die auf diesen Sprachmodellen basierenden KI-Tools häufig besser auf die Bedürfnisse der Nutzenden eingehen können, werden sie öfters Antworten ausgeben, welche den Nutzenden lediglich korrekt erscheinen oder zu deren Weltbild passen (*sycophancy;* Perez et al., 2022).

Darüber hinaus kommt es auch zu fehlerhaften Ausgaben, wenn KI-Tools Informationen «erfinden» oder uns «anlügen».[11] Dieses Phänomen wird als *Halluzinieren* oder *Konfabulieren* bezeichnet (Friedrich et al., 2024; Waldo & Boussard, 2024). Beispielsweise sorgten bereits verschiedenste Gerichtsfälle für Aufsehen, bei welchen von einer Partei – wohl unwissentlich – mit von KI-Tools «halluzinierten» Gerichtsurteilen argumentiert wurde (z. B. Neumeister, 2023). Vor Gericht zeigte sich dann, dass die verwendeten Urteile gar nicht existierten. Den betreffenden Jurist:innen schien entweder nicht bewusst gewesen zu sein, dass KI-Tools auch fehlerhafte Ausgaben machen oder sie überprüften die Ausgaben zu wenig (Klein, 2025). Dieses Fehlverhalten zog teilweise auch rechtliche Sanktionen nach sich, z. B. in Form von Geldstrafen (Merken, 2024).

> «KI-Systeme halluzinieren die ganze Zeit, da sie nichts verstehen und deshalb lediglich halluzinieren können. Häufig sind die ausgegebenen Antworten nur ‚zufällig' richtig, weil die LLMs *[Large Language Models]* mit ausreichend viel Material trainiert wurden.»
>
> —Buck (2025, p. 37)

Auch bei wissenschaftlichen Quellenangaben wird immer wieder von «Halluzinieren» berichtet. Dabei gibt das KI-Tool auf die Frage nach Quellen für eine Aussage authentisch klingende Quellenangaben aus, welche nicht existieren (Agrawal et al., 2024). Besonders problematisch ist hierbei, dass die Quellen meist ein passendes Zitierformat aufweisen (z. B. gemäss den Richtlinien der American Psychological Association, 2020) und sowohl Autor:innen, Titel als auch Publikationswerk zum gewünschten Thema passen. Jedoch haben (a) die Autor:innen nie zusammen eine Publikation geschrieben, (b) der Titel stimmt nicht mit der eigentlichen Publikation der Au-

11 Hier sei nochmals auf die Anthropomorphisierung verwiesen: Ein KI-Tool kann weder lügen, noch erfinden, noch halluzinieren – daher sind diese Begriffe jeweils in Anführungszeichen gesetzt.

tor:innen überein und/oder (c) die Publikation erschien in einer anderen wissenschaftlichen Zeitschrift (Agrawal et al., 2024; Day, 2023). Diese Variationen bleiben auf den ersten Blick möglicherweise unbemerkt, da sie nur bei grosser thematischer Expertise auffallen. Umso wichtiger ist die Verifizierung der Echtheit von Quellen mittels einer eigenen Recherche. Es sei nochmals erwähnt, dass diese unabhängige Prüfung der von KI-Tools ausgegebenen «Wahrheiten» essenziell ist.

Für das «Halluzinieren» eines KI-Tools gibt es mehrere Gründe:

1. Tokens: Die Trainingsdaten werden in Tokens aufgesplittet und Text wird als Aneinanderreihung solcher Tokens generiert. Die Tokens werden dabei Anhand ihrer Zusammenhänge in den Trainingsdaten auf Basis von Wahrscheinlichkeiten ausgewählt (Wolfram, 2023). Die damit generierten Texte sind zwar häufig inhaltlich korrekt, aber längst nicht immer.
2. *Temperature:* Zwar ist immer definiert, welches Token am wahrscheinlichsten als nächstes generiert werden sollte. Aufgrund der *temperature* wird jedoch nicht immer das Token mit der höchsten Wahrscheinlichkeit gewählt. Dadurch wird der Text zwar sprachlich variabler, aber auch inhaltlich variabler – und fehlerhafter (Oertner, 2024).
3. Fehlendes Textverständnis: KI-Tools können nicht überprüfen, ob der generierte Text korrekt ist. Zwar verfügen viele KI-Tools mittlerweile über Internetzugriff, was die Korrektheit und Aktualität der Informationen erhöht. Da sie aber weder ihre eigene Ausgabe noch die Inhalte von Internetquellen «verstehen», ist es ihnen nicht möglich, eigene Fehler zu erkennen (Buck, 2025).

Daneben haben KI-Tools eine weitere Eigenheit, welche das «Halluzinieren» begünstigt: Sie sind grundsätzlich darauf ausgelegt, immer eine Ausgabe zu generieren – unabhängig davon, ob dafür eine Basis in Form von «Hintergrundwissen» vorhanden ist oder nicht (Oertner, 2024).[12] Auch bei fehlender oder limitierter Datenbasis – wie sie z. B. bei spezifischen Fachthemen öfters auftreten wird – gibt ein KI-Tool somit in der Regel eine Antwort aus. Diese Ausgabe wird dann aufgrund der limitierten Datenbasis mit höherer Wahrscheinlichkeit falsch sein (Waldo & Boussard, 2024). Üblicherweise werden daher Angaben zu spezifischen Fachthemen häufiger falsch sein als Angaben zu allgemeinen Themen, zu welchen eine bessere Datengrundlage vorhanden ist.

12 Ausnahmen für den «Ausgabezwang» gibt es zwar, diese wurden aber alle in Form von Filtern in das KI-Tool einprogrammiert (Oertner, 2024).

Verzerrungen *(biases):*

> «Sie *[KI-Textausgaben]* können vorurteilsbehaftet, stereotyp, abergläubisch, weltanschaulich gefärbt oder politisch radikal sein und gründen sich auf Privatmeinungen, unseriöse Berichterstattung, Gossip, Werbung, kommerziell und politisch motivierte Desinformation, Propaganda und Fake News.»
>
> —Oertner (2024, p. 280)

Die Ausgaben von KI-Tools beinhalten neben inhaltlichen Fehlern auch (u. a.) *politische* und *soziale Verzerrungen* (Buck, 2025; Navigli et al., 2023). Aufgrund der anglo-amerikanischen Dominanz bei Anbietern von KI-Tools kommen westliche, englischsprachige Weltanschauungen häufiger in Ausgaben dieser Tools vor. Auch hier liegt die Begründung v. a. in den *Trainingsdaten,* genauer gesagt in deren Sammlung, Vorverarbeitung und Auswahl (Navigli et al., 2023; Paullada et al., 2021). Gewisse Textarten oder Wissensbereiche sind aufgrund ihrer besseren *Verfügbarkeit* – z. B. durch freie Zugänglichkeit im Internet – überrepräsentiert, ebenso wie bestimmte Sprachen und Kulturen. Zudem werden die als Datengrundlage ausgewählten Webseiten oder andere Textkorpora von Personen mit bestimmten Weltanschauungen unterhalten und gepflegt, was ebenfalls zu Verzerrungen in den enthaltenen Informationen führt. So sind z. B. Beitragende auf Wikipedia überwiegend junge und pensionierte Männer (Navigli et al., 2023). Darüber hinaus kann auch die *Vorverarbeitung* und die endgültige *Auswahl* des Trainingsmaterials durch die Entscheidungsträger:innen verzerrt sein (Navigli et al., 2023). Als Folge davon weisen KI-Tools z. B. auch politische Verzerrungen auf, meist sind die Tools eher links orientiert, v. a. bei polarisierenden Themen (K. Yang et al., 2025). Neben dem gewählten KI-Tool hat auch die Nutzungssprache einen Einfluss auf die politische Färbung der Ausgabe (Cahyawijaya et al., 2025; Steinert & Kazenwadel, 2025).

Soziale Verzerrungen bei KI-Tools zeigen sich u. a. hinsichtlich des biologischen Geschlechts, des Alters, der Herkunft und der sexuellen Orientierung (für eine Übersicht, siehe Navigli et al., 2023). Verzerrungen hinsichtlich des Geschlechts sind z. B. bei Übersetzungen von geschlechtsneutralen Begriffen durch KI-Tools möglich, wenn etwa der englische Begriff «nurse» mit der weiblichen Form «Krankenschwester», der Begriff «doctor» hingegen mit der männlichen Form «Arzt» übersetzt wird (Navigli et al., 2023). Die Qualität der Trainingsdaten ist jedoch nicht die einzige Begründung für

solche sozialen Verzerrungen. KI-Tools geben teilweise lediglich tatsächlich existierende soziale Ungleichheiten in Form von Vorurteilen oder Stereotypen wieder, welche sich in den Trainingsdaten als «Spiegel der Gesellschaft» finden (Deutscher Ethikrat, 2023).

Um solche politischen und sozialen Verzerrungen verhindern, sind in KI-Tools oftmals Filter einprogrammiert.[13] Diese Filter greifen jedoch nicht immer oder werden durch sog. «Jailbreaks» umgangen. Umgekehrt greifen Filter teilweise zu weit und KI-Tools antworten als Folge überkorrekt oder geben gar keine Antwort aus, obwohl es sich tatsächlich um eine harmlose Anfrage handelt (Oertner, 2024). Eine einprogrammierte Überkorrektur zeigte sich z. B. beim KI-Bildgenerator Gemini von Google, welcher im Februar 2024 plötzlich ethnisch diverse Bilder u. a. der – historisch ausschliesslich weissen – Gründervätern der USA ausgab. Als Folge der ausgelösten öffentlichen Kritik stoppte Google daraufhin sogar zwischenzeitlich die Möglichkeit, mit Gemini Bilder von Menschen zu generieren (Milmo, 2024). Aufgrund der beschriebenen Unzulänglichkeiten von Filtern stellt sich auch die Frage, ob vorhandene Verzerrungen oder zur Vorbeugung dieser Verzerrungen programmierte, aber fehlgeleitete Filter das «kleinere Übel» darstellen (Ferrara, 2023b).

Fehlende Originalität und inhaltliche Tiefe:

> "Some might say that the output of large language models doesn't look all that different from a human writer's first draft, but, again, I think this is a superficial resemblance. Your first draft isn't an unoriginal idea expressed clearly; it's an original idea expressed poorly, and it is accompanied by your amorphous dissatisfaction, your awareness of the distance between what it says and what you want it to say."
>
> (Manche behaupten vielleicht, dass sich die Ergebnisse von Sprachmodellen nicht wesentlich vom ersten Entwurf menschlicher Autor:innen unterscheiden, aber auch hier handelt es sich meiner Meinung nach nur um eine oberflächliche Ähnlichkeit. Ihr erster Entwurf ist keine klar formulierte, aber unoriginelle Idee; es ist eine originelle Idee, die schlecht ausgedrückt ist, begleitet von Ihrer diffusen Unzufriedenheit und Ihrem Bewusstsein für die Diskrepanz zwischen dem, was sie aussagt, und dem, was Sie eigentlich ausdrücken wollten.)
>
> —Chiang (2023, para. 24)

13 Es gibt aber auch KI-Tools, welche damit werben, ehrlich und nicht politisch korrekt zu sein, z. B. Grok von Elon Musks KI-Unternehmen xAI (M. Yang, 2025).

Bei von KI-Tools generierten Texten kommt es vor, dass diese im Vergleich zu von Menschen geschriebenen Texten weniger originelle Inhalte und überzeugende Argumente aufweisen (Boussioux et al., 2024; Niloy et al., 2024; Stadler et al., 2024). Bereits im Kapitel 2.1 Phasenmodell des Schreibprozesses wurde darauf eingegangen, dass Schreiben auch dazu dient, die eigenen Gedanken zu ordnen *(writing is thinking on paper)*. Werden Texte bereits in einem ersten Schritt von KI-Tools generiert, gehen womöglich inhaltlich relevante Gedankengänge verloren oder argumentative Lücken bleiben unerkannt, was oberflächliche Texte zur Folge hat (Buck, 2025).

> «Die enervierendste Fehlleistung des Chatbots besteht sicherlich darin, weitschweifige, substanzlose Satzhülsen zu erzeugen, triviale Wahrheiten ohne Erkenntnisgewinn zu verbreiten und sich in endlosen Wiederholungen zu ergehen.»
>
> —Oertner (2024, p. 287)

Die teilweise fehlende inhaltliche Varianz in den Texten («Satzhülsen») wird dadurch begründet, dass KI-Tools Informationen nicht gewichten können und somit nur die am häufigsten in den Trainingsdaten vorkommenden Zusammenhänge wiedergegeben werden. Dadurch wird auch klar, weshalb KI-Tools oftmals nicht auf den spezifischen Kontext einer Anfrage eingehen können (Cheung, 2024). Sollen z. B. Prozesse in einem Unternehmen analysiert oder beschrieben werden, hat das KI-Tool darüber schlicht zu wenig Wissen und macht dann nur vergleichsweise allgemeine Aussagen mit hoher «Flughöhe». Aber selbst wenn ein KI-Tool durch geeignetes Prompting viele Ideen generiert, werden diese nicht gewichtet und es braucht menschliches Zutun, denn ein KI-Tool kann nicht identifizieren, welches die innovativste bzw. für die Anforderung der Nutzenden am besten passende Idee ist (Becker, 2024; Dwivedi et al., 2023).

Ausgaben von KI-Tools enthalten zum Teil auch nur unvollständige Informationen. Die ausgegebenen Inhalte sind zwar korrekt, aber durch die Unvollständigkeit kann die Qualität des Textes vermindert werden. Insbesondere wenn ein KI-Tool Listen oder Tabellen ausgibt, führt dies allenfalls zum unzutreffenden Eindruck, dass alle relevanten Informationen zu einem Thema vorhanden sind. Dadurch wiegen sich Nutzende womöglich in falscher Sicherheit und unternehmen keine weiteren Schritte – wie z. B. eine eigene Recherche durchzuführen oder eigene Gedanken einzubringen –, um einem Text eine angemessene inhaltliche Tiefe zu verleihen (Oertner, 2024).

Veränderung der Sprache:

> "While AI provides valuable support, remember that the heart of any great book is the unique touch that only you, the writer, can provide."
> (Auch wenn KI eine wertvolle Unterstützung ist, vergessen Sie nicht, dass das Herzstück eines jeden guten Buches die einzigartige Note ist, die nur Sie als Schreibende verleihen können.)
>
> —Kosberg (2024, p. 57)

KI-generierte Texte werden nicht nur verschiedentlich als *inhaltlich* oberflächlich beschrieben, sondern auch z. B. als *sprachlich* «variationsarm», «steril», «starr», «floskelhaft», «leer» und «distanziert» (Jiang & Hyland, 2025, p. 375; Markey et al., 2024, p. 571; Schicker & Akbulut, 2023, p. 191). Darüber hinaus wird teilweise der fehlende emotionale Tiefgang solcher Texte bemängelt (Barrot, 2023; C. Wang et al., 2024). Diese beiden Aspekte sind eng miteinander verwandt und hängen mit der Funktionsweise von KI-Tools zusammen: Bestimmte, häufig verwendete Wörter und Wortfolgen werden häufiger generiert, was die sprachliche und emotionale Variation in den ausgegebenen Texten einschränkt.

Zusammengefasst wird hier auch von einer *Verwässerung des eigenen Schreibstils* gesprochen (Barrot, 2023; Monckton, 2025; C. Wang et al., 2024). Je stärker der Text durch KI-Tools verarbeitet wurde, desto eher ist eine solche Verwässerung die Folge. Auch kollektiv kommt es durch die Nutzung von KI-Tools zu einer veränderten Sprachnutzung. Erste Studien deuten bereits darauf hin, dass von KI-Tools präferierte Wörter mittlerweile häufiger bei der wissenschaftlichen Kommunikation genutzt werden als vor dem Aufkommen dieser Tools (Kobak et al., 2025; Yakura et al., 2024).

Es dürfte kaum überraschen, dass eine Verwässerung des eigenen Schreibstils bei generierten Texten auftritt. Zwar ist es durch gezieltes Prompting sowie spezifische Überarbeitungen möglich, einen Text zu erstellen, der von Schreibenden als eigenes Produkt wahrgenommen wird. Ob jedoch auch Aussenstehende den Stil der Schreibenden identifizieren können, ist dadurch nicht garantiert (Mikros, 2025). Gleiches trifft aber ggf. sogar zu, wenn Texte von KI-Tools lediglich überarbeitet oder korrekturgelesen werden. Bei solchen Textbearbeitungen haben KI-Tools teilweise den «Hang» zur Überkorrektur, wodurch dem eigenen Schreiben inhärente sprachliche Feinheiten verlorengehen (Baron, 2024).

Sprachliche Fehler und Inkonsistenzen:

Auch wenn die Texte sprachlich ausgereift klingen, darf dies nicht darüber hinwegtäuschen, dass KI-Tools Texte mit sprachlichen Fehlern und Inkonsistenzen ausgeben. Dazu zählen z. B. grammatikalische Fehler (Borji, 2023), fehlende oder doppelte Verneinungen (Oertner, 2024), unnötige sprachliche Wiederholungen (Schicker & Akbulut, 2023) oder eine inkonsistente Verwendung von Begrifflichkeiten (Baron, 2024; Morrison, 2023).

Über die Textgenerierung hinaus bestehen diese Probleme auch bei der Textüberarbeitung *(editing)* und dem Korrekturlesen *(proofreading)*. KI-Tools können zwar bei diesen Aufgaben helfen. Dabei muss aber berücksichtigt werden, dass diese Tools die Texte, welche sie bearbeiten sollen, nicht verstehen. Daher ist es wenig überraschend, dass genutzte KI-Tools hierbei sprachliche Fehler und Inkonsistenzen im zu bearbeitenden Text «übersehen» oder sogar neue Fehler und Inkonsistenzen «einbauen» (Baron, 2024; Monckton, 2025; Montgomerie, 2024).

Transparenz

Chancen

Insbesondere bei wissenschaftlichen Texten ist es ein zentrales Element, dass Autor:innen bezüglich deren Entstehung transparent sind: Auf welchen theoretischen und empirischen Quellen basieren die geäusserten Annahmen? Welche Methodik und welche Daten wurden verwendet? Wie kommen die Autor:innen zu ihren Schlussfolgerungen? KI-Tools unterstützen bei der Sicherstellung der Transparenz, indem sie z. B. bei der Literaturrecherche helfen, das Verständnis von Quellen erleichtern oder die Zuordnung von Aussagen zu dazugehörigen Quellen verbessern (Gredel et al., 2024; Huang & Tan, 2023). Wichtig ist hier zu erwähnen, dass die Autor:innen selbst dafür verantwortlich sind, dass die Entstehung der Texte transparent dokumentiert ist. KI-Tools haben hier lediglich unterstützende Funktion.

Risiken

Die *mangelnde Transparenz der Ausgabe* von KI-Tools muss als Risiko bezeichnet werden. Für die Nutzenden bleibt unklar, woher die Informationen stammen, welche vom KI-Tool ausgegeben werden (Arnold, 2024). Dabei ist insbesondere problematisch, dass die *Datengrundlage* in der Form der Trainingsdaten selektiv und damit verzerrt ist. Zum einen werden frei verfügbare Inhalte wie Texte auf Webseiten oder öffentliche Bücher aus dem Internet gesammelt. Zum anderen gehen Anbieter von KI-Tools immer häufiger Partnerschaften mit Unternehmen ein, um deren Textmaterial für die Trainingsdaten nutzen zu können. So kooperiert z. B. OpenAI mit dem Axel-Springer-Verlag, welcher mit Publikationen wie der Bild-Zeitung nicht unbedingt für seriösen Journalismus und ausgewogene Berichterstattung bekannt ist (Troendle, 2023). Solche Partnerschaften führen in vielen Fällen zu einer Verzerrung der in den Trainingsdaten vorhandenen Informationen in eine bestimmte Richtung, was für die Nutzenden der KI-Tools ohne Hintergrundwissen nicht erkennbar ist. Die Ausgaben müssen somit als das Ergebnis einer nicht nachvollziehbaren Verrechnung einer einseitigen Selektion von Primär-, Sekundär- und Tertiärliteratur bezeichnet werden.

Dazu kommt das in diesem Zusammenhang häufig genannte *Black Box-Problem:* Von aussen ist im Einzelfall kaum nachvollziehbar, wie die eingegebenen Daten zu Ausgaben verarbeitet werden.[14] Wie wurden die Daten vorverarbeitet? Wie wurden die Sprachmodelle trainiert? Es besteht dabei Trade-off zwischen Zuverlässigkeit und Transparenz: Je komplexer die zugrundeliegenden Sprachmodelle, desto genauer sind KI-Tools, aber desto weniger transparent ist auch ihre Funktionsweise (Buck, 2025).

Neben der Datengrundlage erschwert auch die *Funktionsweise* von KI-Tools die Transparenz. Eine identische Eingabe in ein KI-Tool führt in der Regel nicht immer zu derselben Ausgabe. Besonders im wissenschaftlichen Kontext, wo der Transparenz eine besonders wichtige Stellung zukommt, ist dieser Umstand problematisch. Wissenschaftliche Methodik sollte von Forschenden immer so beschrieben werden, dass sie von Aussenstehenden wiederholt werden kann und diese zu denselben Schlüssen kommen wie die Forschenden. Ist dies nicht gegeben, beeinträchtigt dies die Glaubwürdig-

14 Der Begriff der Black Box ist in diesem Zusammenhang umstritten, denn die Trainingsdaten und die grundsätzliche Funktionsweise der KI-Tools sind zumindest teilweise bekannt (Buck, 2025).

keit von wissenschaftlichen Ergebnissen und damit den wissenschaftlichen Fortschritt als Ganzes (Arnold, 2024).

Darüber hinaus kann ein KI-Tool auch nicht selbst erklären, wie es zu einem Ergebnis gekommen ist, was die Transparenz ebenfalls einschränkt (Dwivedi et al., 2023). Zwar geben KI-Tools Quellen zu generierten Informationen aus, wenn sie danach gefragt werden. Diese Quellen existieren teilweise auch tatsächlich und passen zum Thema des generierten Textes. Allerdings werden die Quellen nicht auf korrekte Art und Weise dem generierten Text zugeordnet, sondern lediglich aufgrund einer anschliessenden Internetrecherche durch das KI-Tool hinzugefügt (Oertner, 2024). Somit basieren die Aussagen im generierten Text nicht auf den ausgegebenen Quellen und stimmen nicht zwingend mit den Aussagen in den Quellen überein – bzw. wenn doch, dann nur per Zufall.

Mittlerweile gibt es zwar neue Funktionen in KI-Tools, bei welchen tatsächlich eine aktive Internetrecherche durchgeführt und auf deren Basis eine Ausgabe generiert wird. Diese sind jedoch meist nur bei kostenpflichtigen Versionen uneingeschränkt verfügbar und auch hier ist aufgrund der Funktionsweise nicht ausgeschlossen, dass fehlerhafte Informationen ausgegeben werden – d. h. die Aussagen im generierten Text nicht mit den Aussagen in der genannten Quelle übereinstimmen. Zudem besteht auch bei diesem Vorgehen ein Problem bzgl. der Transparenz. Anders als z. B. bei einer Google-Suche wird nicht das gesamte Internet durchsucht, sondern lediglich vom Anbieter des KI-Tools definierte Webseiten – diese Einschränkung ist jedoch für Nutzende nicht ohne weiteres erkennbar (Heikkilä & Honan, 2024).

Kompetenzerwerb

Chancen

> "By engaging with AI through a series of interactive prompts, individuals can receive input that is comprehensible and slightly beyond their current competence."
>
> (Durch interaktive Prompts an die KI können Nutzende Input erhalten, der für sie verständlich ist und leicht über ihre derzeitige Kompetenz hinausgeht.)
>
> —Ingley und Pack (2023, p. 785)

KI-Tools können Schreibende beim Lernen unterstützen (*upskilling;* Rafner et al., 2021). Sowohl bei der Aneignung von Wissen als auch bei der Aneignung von (Schreib-)Kompetenzen ist KI bei richtiger Anwendung ein nützliches Werkzeug (Deng et al., 2025; Kasneci et al., 2023; J. Wang & Fan, 2025).

1. Aneignung von Wissen und Kompetenzen (allgemein)
   - Schrittweise Erweiterung des eigenen Wissens: Durch eine wiederholte Interaktion mit einem KI-Tool kann das eigene Wissen schrittweise erweitert werden. KI-Tools helfen dabei, unbekannte Themenfelder verständlich zu machen – indem z. B. schwer verständliche Texte oder Fachbegriffe auf unterschiedliche Weise wiedergegeben werden (Farrokhnia et al., 2024) – und unterstützen durch die Ausgabe von relevanten Hintergrundinformationen und unterschiedliche Perspektiven zu einem Thema (Deng et al., 2025).
   - Unmittelbares, personalisiertes Lernen: Die Interaktion mit dem KI-Tool erlaubt gezielte Nachfragen bei Unklarheiten und eine sofortige, individuelle Antwortausgabe durch das KI-Tool. Diese Art von Feedback ermöglicht personalisiertes Lernen (Farrokhnia et al., 2024) und fördert dadurch die Lernmotivation (Chang et al., 2025).
   - Zugang zu breitem Fachwissen: Das von KI-Tools «abrufbare» Wissen und das damit verbundene Potential ist riesig: ChatGPT hat z. B. ohne spezielles Training bereits verschiedenste Abschlussprüfungen oder zumindest Teile davon bestanden, z. B. in der Medizin (Kung et al., 2023; Suwała et al., 2024) und in den Rechtswissenschaften (Katz et al., 2023).

2. Aneignung von Schreibkompetenz

> "GenAI chatbots can offer insights into your writing in ways that, in some cases, a writing tutor cannot."
>
> (Generative KI-Chatbots können Ihnen Einblicke in Ihr Schreiben geben, welche Ihnen ein Schreibcoach in manchen Fällen nicht bieten kann.)
>
> —Dobrin (2023, p. 21)

   - Verbesserung des Sprachverständnisses: KI-Tools ermöglichen es, z. B. sprachliche Fehler zu ermitteln und direkt zu korrigieren oder passendere Begriffe und Formulierungen zu identifizieren, wodurch das Sprachverständnis langfristig verbessert wird (Kasneci et al., 2023; Z. Lin, 2024b).

- Verbesserung von Formulierung und Argumentation: KI-Tools ermöglichen es, Feedback zum Text aus unterschiedlichen Perspektiven zu erhalten – z. B. aus Sicht eines Laien, einer Lehrperson oder eines kritischen Reviewers von wissenschaftlichen Zeitschriftenartikeln. Dadurch wird das Verständnis darüber gefördert, wie der Text verstanden – bzw. nicht verstanden – wird und wo mögliche inhaltliche Schwachpunkte liegen. Langfristig werden dadurch Kompetenzen aufgebaut, wie Texte für verschiedene Zielgruppen verständlich formuliert und überzeugend argumentiert werden.

Risiken

> "While generative AI tools such as ChatGPT can make tasks significantly easier for humans, they come with the risk of deteriorating our ability to effectively learn some of the skills required to solve these tasks."
>
> (Zwar können generative KI-Tools wie ChatGPT Aufgaben für Menschen erheblich erleichtern, sie gehen jedoch mit dem Risiko einher, dass unsere Fähigkeit zum tatsächlichen Erwerb einiger zur Erfüllung dieser Aufgaben notwendigen Kompetenzen beeinträchtigt wird.)
>
> —Bastani et al. (2025, Chapter Discussion)

Individueller Kompetenzverlust:

Ein wesentliches Risiko bei der Nutzung von KI-Tools liegt in einem Verlust von Kompetenzen, dem sog. *deskilling* (Deutscher Ethikrat, 2023; Rafner et al., 2021; Reinmann, 2023). Im Zusammenhang mit dem Schreiben sind als relevante betroffene Kompetenzen insbesondere das (a) kritische Denken *(critical thinking),* die (b) Informationskompetenz (d. h. Informationen suchen und bewerten können) und die (c) Schreibkompetenz (d. h. sich schriftlich ausdrücken können, inkl. Grammatik und Rechtschreibung) zu nennen (Albrecht, 2023). Diese drei Kompetenzen können durch die Nutzung von KI-Tools verringert werden oder verloren gehen – oder bereits deren Aufbau wird behindert (Gerlich, 2025; Heersmink, 2024; Kosmyna et al., 2025).

> "A key irony of automation is that by mechanising routine tasks and leaving exception-handling to the human user, you deprive the user of the routine opportunities to practice their judgement and strengthen their cognitive musculature, leaving them atrophied and unprepared when the exceptions do arise."
>
> (Eine wesentliche Ironie der Automatisierung besteht darin, dass dadurch, dass Routineaufgaben an Maschinen abgegeben und Ausnahmen den menschlichen Nutzenden überlassen werden, diesen die Möglichkeit genommen wird, ihr Urteilsvermögen zu trainieren und ihre kognitive Muskulatur zu stärken, so dass diese verkümmert und unvorbereitet sind, wenn die Ausnahmen auftreten.)
>
> — H.-P. Lee et al. (2025, Chapter 1)

Ein wesentlicher Grund für diesen möglichen Kompetenzverlust besteht darin, dass die Kompetenzen durch die Abgabe von Aufgaben und Teilschritten an KI-Tools weniger oder lediglich oberflächlich geübt werden (Bastani et al., 2025; Kosmyna et al., 2025; Lehmann et al., 2025). Ähnlich wie sich Muskeln abbauen, wenn sie nicht trainiert werden, gehen auch Kompetenzen mit der Zeit verloren, wenn sie aufgrund der Nutzung von KI (zu) wenig beübt werden (Rafner et al., 2021). Wie z. B. beim Liftfahren statt Treppensteigen bildet sich dann lediglich die Illusion, die Höhe selbst zurückgelegt bzw. die Aufgabe selbst bewältigt zu haben.

Ein weiterer Grund besteht im sog. *automation bias:* KI-Tools – und Maschinen im Allgemeinen – werden teilweise für «allwissend, unfehlbar und überlegen» (Oertner, 2024, p. 263) gehalten. Infolgedessen werden KI-generierte Ausgaben unkritisch akzeptiert und auf deren Basis gehandelt und entschieden (Albrecht, 2023; Deutscher Ethikrat, 2023).[15] Dieses Verhalten wird auch als *overreliance* (übergrosses Vertrauen) bezeichnet. Ein solches übergrosses Vertrauen in KI-Tools zeigt sich z. B., wenn Studierende selbst in ihrem eigenen Fachgebiet KI-Tools mehr trauen als dem eigenen Fachwissen und dadurch auch fehlerhafte Ausgaben von KI-Tools übernehmen (Krupp et al., 2024). Die von KI-Tools ausgegebenen Texte scheinen eine solch hohe Qualität zu haben, dass Nutzende sich die Mühe lieber sparen und gleich die generierte Ausgabe verwenden – ohne zu beachten, dass dadurch die eigenen Kompetenzen verloren gehen.

15 Mit der Anthropomorphisierung gibt es eine zum *automation bias* gegenteilige Haltung zu KI-Tools. Diese Vermenschlichung sollte den *automation bias* eigentlich abschwächen. Stattdessen scheinen jedoch beide Haltungen, welche teilweise parallel existieren, zu mehr Vertrauen in die KI-Tools zu führen (für eine mögliche Erläuterung, siehe Oertner, 2024).

Gesellschaftlicher Kompetenzverlust:

> «Die möglichen gesellschaftlichen Folgen sind ernst und demokratiebedrohend. Die massenhafte Verbreitung von Fehlinformation und Texthülsen hat eine Kontaminierung und Verwässerung des Weltwissens zur Folge, die sich rekursiv beschleunigt, wenn Sprachmodelle Sprachmodelle füttern, direkt über künstliches Trainingsmaterial oder indirekt über KI-generierten Internet-Content.»
>
> —Oertner (2024, pp. 291–292)

Wenn Kompetenzen gar nicht mehr erworben werden, kann – zumindest längerfristig – auch von *deskilling* auf gesellschaftlicher Ebene gesprochen werden (Reinmann, 2023). So dienen z. B. schriftliche Arbeiten im Hochschulkontext dazu, dass Studierende u. a. lernen, eine Forschungsfrage aus der wissenschaftlichen Literatur herzuleiten (Informationskompetenz), zu formulieren und zu beantworten (Methodenkompetenz). Darüber hinaus dienen diese Arbeiten dem Erwerb und der Förderung der Fähigkeit, den Schreibprozess zur Entwicklung von Gedanken zu einem Thema zu nutzen (kritisches Denken) und diese schriftlich zu kommunizieren bzw. sich generell schriftlich ausdrücken zu können (Schreibkompetenz). Werden mehrheitlich oder ausschliesslich KI-Tools zur Entwicklung und Formulierung von Ideen genutzt, so wird der Kompetenzerwerb dadurch bestenfalls beschränkt und schlimmstenfalls verunmöglicht.

> "One of the most pressing concerns is the potential erosion of our critical thinking and reasoning abilities. If we blindly trust the output of … [AI systems] without questioning or scrutinizing it, we risk losing our ability to think independently and form our own judgments."
>
> (Eines der drängendsten Probleme ist die mögliche Erosion unserer Fähigkeiten zum kritischen Denken und Schlussfolgern. Wenn wir den Ergebnissen von KI-Systemen blind vertrauen, ohne sie zu hinterfragen oder zu überprüfen, riskieren wir, unsere Fähigkeit zum unabhängigen Denken und zur Bildung eigener Urteile zu verlieren.)
>
> —Chiriatti et al. (2024, p. 1829)

Bei weit fortgeschrittenem *deskilling* auf gesellschaftlicher Ebene wäre eine regelrechte Abhängigkeit von KI-Tools die Folge (Bendel, 2024). Funktioniert ein KI-Tool daraufhin nicht wie es sollte, kann es nicht ordnungsgemäss überwacht und korrektiv eingegriffen werden, weil die dafür notwen-

digen Kompetenzen nicht mehr vorhanden sind (Deutscher Ethikrat, 2023; Rafner et al., 2021).

Risiko für Wissenschaft und Informationsgesellschaft:

KI-Tools stellen auch ein Risiko für die Wissenschaft und die Informationsgesellschaft dar. Neben dem *deskilling* haben hierbei auch bereits genannte Faktoren wie die unbeabsichtigte (bzw. fahrlässige) oder absichtliche Verbreitung von fehlerhaften Informationen (Chen & Shu, 2024), mangelnde Transparenz sowie die Verwässerung des Schreibstils – und damit verringerte Authentizität – einen Einfluss.

Die weit verbreitete Nutzung von KI-Tools führt mit hoher Wahrscheinlichkeit dazu, dass die Anzahl erstellter wissenschaftlicher Texte stark zunimmt, ohne dass deren Qualität in vergleichbarer Weise steigt (Naddaf, 2025). Die Folge wären mehr oberflächliche und/oder qualitativ minderwertige – bzw. lediglich wissenschaftlich klingende – Texte (Albrecht, 2023; Bendel, 2024). Durch diese Mehrproduktion könnte das wissenschaftliche *peer review* System stark belastet oder sogar überlastet werden (Bendel, 2024). Bei einer Überlastung des *peer review* Systems würden ggf. immer häufiger KI-Tools zur Erstellung von *peer reviews* genutzt und damit schlimmstenfalls ein System geschaffen, in welchem KI-Tools sowohl wissenschaftliche Texte schreiben als auch deren Qualität überprüfen. Gleichzeitig würden durch den vorherrschenden Publikationsdruck so zudem zukünftig wahrscheinlich immer mehr qualitativ minderwertige Artikel in sog. *predatory journals* veröffentlicht (Kim, 2023). Wenn es langfristig nicht mehr möglich ist, qualitativ hochwertige Texte von KI-generiertem *garbage* zu unterschieden, könnte das Vertrauen in die Wissenschaft und in verbreitete Informationen verloren gehen (Oertner, 2024).

Zwar lässt sich durch eine umsichtige Nutzung sowie Regulierung von KI-Tools in der Wissenschaft diesen Risiken teilweise vorbeugen. Selbst dann besteht jedoch das Risiko für Verständnisillusionen *(illusions of understanding),* also das durch KI-Tools geförderte Gefühl, mehr über die Welt zu verstehen, als eigentlich der Fall ist (Messeri & Crockett, 2024). Als Folge einer solchen Verständnisillusion wäre die Forschung einseitiger, weniger innovativ und fehleranfälliger (Messeri & Crockett, 2024). Insbesondere seltenes Spezialwissen wird bei KI-generierten Ausgaben öfters vernachlässigt, weil es in den Trainingsdaten weniger vorkommt, was langfristig sogar zu einem Verlust von Wissen *(knowledge collapse)* führen kann (Peterson, 2025).

Handlungsfähigkeit und Motivation beim Schreiben

Chancen

> "At its core, AI is a creativity booster. Imagine sitting and brainstorming with a coach who's read a library's worth of books and can generate ideas at lightning speed."
>
> (Im Grunde ist KI ein Kreativitätsverstärker. Stellen Sie sich vor, Sie sitzen bei einem Brainstorming mit einem Coach zusammen, der eine ganze Bibliothek von Büchern gelesen hat und blitzschnell Ideen entwickeln kann.)
>
> —Kosberg (2024, p. 54)

Über die bereits in Kapitel 2.4 Umgang mit Schreibblockaden genannten Ansätze hinaus sind KI-Tools auch bei der Überwindung von Schreibblockaden hilfreich.

- Denkanstoss: Die Nutzung von KI-Tools bei Schreibblockaden bedeutet nicht, einen Text von einem KI-Tool generieren und unverändert übernehmen zu lassen. Stattdessen kann ein KI-Tool auch lediglich als Denkanstoss dienen, mithilfe dessen der Einstieg in einen Text erleichtert wird (Kosberg, 2024).
- Strukturierung: Auch bei der Strukturierung von Ideen und einzelnen Abschnitten bieten KI-Tools Unterstützung, sodass der Schreibprozess flüssiger gestaltet wird oder Schreibblockaden sogar verhindert werden (Kosberg, 2024).
- Verringerung von Angst: Teilweise geht mit Schreibblockaden die Angst einher, einen unfertigen Text mit einer anderen Person zu teilen. Diese Angst lässt sich durch die Nutzung von KI-Tools verringern. Autor:innen erhalten dann zuerst Rückmeldung durch das KI-Tool und können den Text überarbeiten, ehe sie ihn mit anderen Personen teilen (Buck, 2025).

Durch die oben genannten Chancen von KI beim Schreiben werden zudem die Freude und Motivation beim Schreiben gestärkt. Faktoren wie höhere Effizienz, höhere Textqualität, Kompetenzerwerb sowie ausbleibende Schreibblockaden führen in der Regel zu einer höheren wahrgenommenen Handlungsfähigkeit und tragen so zu einem positiven Erleben des Schreibens bei. Auch durch direktes, personalisiertes Feedback sowie die wahrgenommene Unterstützung durch KI-Tools wird die Freude und Motivation beim Schreiben in vielen Fällen erhöht (Deng et al., 2025; Meyer et al., 2024; Teng, 2024).

Risiken

Schreibblockaden könnten künftig zunehmen, weil aufgrund der allgemein steigenden Textqualität durch KI-Unterstützung die – tatsächlichen oder wahrgenommenen – Ansprüche an Texte höher werden (Bräuer et al., 2023). Darüber hinaus ist eine Folge des *deskilling* auch eine Einschränkung der eigenen Handlungsfähigkeit, weil Autor:innen sich nicht mehr in der Lage fühlen, Texte selbstständig zu schreiben. Dadurch entsteht ein Gefühl der «Fremdbestimmtheit» und eine Aushöhlung von Autorschaft und Verantwortung für geschriebene Texte (Deutscher Ethikrat, 2023; Reinmann, 2023). Mithilfe von KI-Tools geschriebene Texte werden so allenfalls nicht mehr als selbst verfasst wahrgenommen (Kosmyna et al., 2025).

Auch die Schreibmotivation kann durch die Nutzung von KI-Tools verringert werden. Die Wahrnehmung, dass «auf Knopfdruck» ein qualitativ hochwertiger Text entsteht, führt allenfalls dazu, dass der Mehrwert der eigenen (Denk-)Arbeit nicht mehr gesehen wird. Dadurch sinkt die Bereitschaft, Zeit und Ressourcen in das Schreiben zu investieren (*law of diminishing returns).* Umgekehrt scheint es wahrscheinlich, dass auch nicht zufriedenstellende Ausgaben von KI-Tools die Motivation beim Schreiben verringern, indem diese den Schreibprozess behindern (Montes & Khojah, 2025; T.-C. Yang et al., 2025).

Fazit

> "Success in creating AI could be the biggest event in the history of our civilization. But it could also be the last – unless we learn how to avoid the risks."
>
> (Der Erfolg bei der Entwicklung künstlicher Intelligenz könnte das grösste Ereignis in der Geschichte unserer Zivilisation sein. Es könnte aber auch das Letzte sein – es sei denn, wir lernen, wie wir die Risiken vermeiden können.)
>
> —Stephen Hawking (1942–2018)

Die Nutzung von KI-Tools für das Schreiben birgt sowohl Chancen als auch Risiken. Ein zentrales Problem sind dabei auch falsche Vorstellungen der Funktionsweise von KI-Tools und damit einhergehend eine Fehleinschätzung bzw. Überschätzung derer Möglichkeiten / «Fähigkeiten» und Einsatz-

bereiche.[16] Dadurch kann es zu einer falschen Nutzung dieser Tools kommen, was sich negativ auf das Schreiben auswirkt. Um dies zu verhindern, wird im folgenden Kapitel 7.3 Schreiben mit KI ein verantwortungsvoller Umgang mit KI-Tools beim Schreiben vorgestellt.

16 Bei gewissen Aspekten besteht hier jedoch auch in der Forschung noch Uneinigkeit, siehe z. B. Shojaee et al. (2025) und Varela et al. (2025).

7.3 Schreiben mit KI

Inhaltsübersicht

> "Generative AI pretends that there can be writing without writers, which is as nonsensical as suggesting there can be swimming without swimmers, or breathing without breathers."
>
> (Generative KI tut so, als könne es Schreiben ohne Schreibende geben, was genauso unsinnig ist wie die Annahme, dass es Schwimmen ohne Schwimmende oder Atmen ohne Atmende geben könne.)
>
> —Morrison (2023, p. 160)

Die Risiken beim Schreiben mit KI sollen nicht darüber hinwegtäuschen, dass die Nutzung von KI-Tools enormes Potential birgt. Zudem ist es weder möglich noch sinnvoll, Fortschritt aufzuhalten. Fortschritt, v. a. technischer Fortschritt, lässt sich nicht aufhalten. Fortschritt ist allerdings nicht mit Innovation gleichzusetzen. Fortschritt setzt voraus, dass eine innovative Idee oder Technologie in einer Verbesserung der institutionellen Rahmenbedingungen und des gesellschaftlichen Lebens resultiert. Keinesfalls darf ein innovatives Produkt Schreibende daran hindern, Wissen und Kompetenzen zu erwerben und sie dadurch ihrer persönlichen Entwicklung berauben.

In diesem Kapitel wird daher beschrieben, wie mit KI-Tools verantwortungsvoll umgegangen werden kann. Dabei wird zuerst auf die Grundhal-

tung von KI als «Copiloten» eingegangen. Danach werden Tipps und Checklisten zum Prompting vorgestellt. Zuletzt werden verschiedenste Anwendungsbereiche von KI-Tools beim Schreiben erörtert.

Grundhaltung: KI als «Copilot»

> "Treat AI like an infinitely patient new coworker who forgets everything you tell them each new conversation, one that comes highly recommended but whose actual abilities are not that clear."
> (Behandeln Sie die KI wie eine unendlich geduldige neue Mitarbeitende, welche bei jedem neuen Gespräch alles vergisst, was Sie ihr sagen. Eine Mitarbeitende, welche wärmstens empfohlen wurde, deren tatsächliche Fähigkeiten aber nicht so klar sind.)
>
> —Mollick (2024b, Section Good Enough Task Prompting)

KI-Tools sollten lediglich die Funktion einer Assistenz übernehmen. Wie der Name *copilot* (d. h. erste Offizier:in) des KI-Tools von Microsoft impliziert, kommt den Schreibenden die Rolle *captain* (d. h. Kommandant:in) zu. Die Schreibenden müssen die Arbeit des KI-Tools beurteilen können und tragen als Urheber:innen die alleinige Verantwortung für das Endergebnis. In diesem Zusammenhang wird auch von *AI Leadership* (Buck & Wessels, 2025) gesprochen.

> «Sie als schreibende Person sind die Führungskraft eines Teams von verschiedenen KI-Tools, die als Ihre Assistenten tätig sind. Diese Assistenten führen Sie als *AI-Leader* mit dem Ziel, das bestmögliche Endprodukt zu erreichen. Als Führungskraft können Sie die letztliche Gestaltungs- und Steuerungshoheit über den Arbeitsprozess aber nicht an Ihre Assistenten abgeben.»
>
> —Buck (2025, p. 75)

Schreibende sind dabei Kommandant:innen bzw. Führungskräfte, welche den Einsatz von KI-Tools als ihre Assistenzen (a) planen, (b) steuern und (c) überprüfen. Das Ziel ist stets ein bestmögliches Endergebnis bei gleichzeitiger Beibehaltung der Kontrolle und Verantwortung.

1. Planung

- Empfehlung: Setzen Sie KI-Tools gezielt ein.
- Erläuterungen:
  - Führungskräfte von KI-Tools erstellen einen Arbeitsplan und verteilen Aufgaben an KI-Tools als deren Assistenzen, ähnlich wie beim Projektmanagement mit menschlichen Assistenzen. Im Zentrum steht die aktive gedankliche Auseinandersetzung mit der Frage, für welche Aufgaben KI-Tools wie eingesetzt werden. Dies erfordert ein gewisses Mass an Klarheit bezüglich der zu erledigenden Aufgaben und ein Wissen über die Möglichkeiten und Grenzen der genutzten KI-Tools.
  - Bereits in dieser Phase wird definiert, anhand welcher Kriterien die Ausgabe der KI-Tools überprüft werden soll. Dabei geht es um eine erste, allgemeine Überprüfung, um sicherzustellen, dass die Ausgabe in die gewünschte Richtung geht: Was benötige ich genau vom KI-Tool für die entsprechende Aufgabe? So kann verhindert werden, dass sich Nutzende mit einer nicht zielführenden Ausgabe des KI-Tools zufriedengeben.

2. Steuerung

- Empfehlung: Steuern Sie den Einsatz von KI-Tools aktiv.
- Erläuterungen:
  - Die Steuerung der KI-Tools erfolgt – zumindest bei chatbasierten KI-Tools – mittels zielgerichtetem, meist wiederholtem Prompting. Der Inhalt des Prompts ergibt sich aus der in der Planungsphase identifizierten Zielvorstellung für die Ausgabe. Je klarer der Prompt hinsichtlich des definierten Ziels formuliert ist, desto eher führt er zur gewünschten Ausgabe.
  - Die Ausgabe kann dann anhand der in der Planungsphase definierten Kriterien überprüft werden. Falls der ursprünglich formulierte Prompt nicht zur gewünschten Ausgabe führt, wird er überarbeitet und/oder durch weitere Prompts ergänzt.

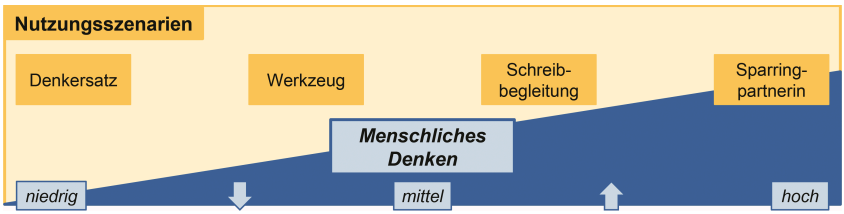
3. Überprüfung

- Empfehlung: Überprüfen Sie die Ausgabe von KI-Tools kritisch.
- Erläuterungen:
  - *Inhaltliche Angaben* von KI-Tools müssen anhand externer Quellen, eigener Expertise und kritischem Denken auf ihre Korrektheit und logische Stringenz überprüft werden (Huang & Tan, 2023; Salvagno et al., 2023; Wan, 2023).
  - Auch eine Überprüfung von allfälligen *Verzerrungen* – z. B. einseitige Gewichtung von Informationen sowie politische oder soziale Verzerrungen – erfolgt in diesem Schritt.
  - Darüber hinaus werden die ausgegebenen *Quellen* überprüft. Es liegt in der Verantwortung der Schreibenden sicherzustellen, Textstellen lediglich mit Quellen zu versehen, welche die im Text enthaltenen Aussagen tatsächlich stützen. Im wissenschaftlichen Kontext – und somit auch im Hochschulkontext – gebietet es die wissenschaftliche Redlichkeit zudem, ausschliesslich Quellen zu zitieren, die eigenhändig konsultiert und verstanden wurden.
  - *Andere Qualitätsaspekte,* wie (a) die Originalität und inhaltlichen Tiefe, (b) die Wahrung des eigenen Schreibstils sowie (c) die sprachliche Korrektheit sollten in dieser Phase zwar ebenfalls geprüft werden, liegen dabei aber – anders als die inhaltliche Korrektheit – im Ermessen der Schreibenden.

Nutzungsszenarien von KI-Tools beim Schreiben

Bei der Nutzung von KI-Tools beim Schreiben lassen sich verschiedene Szenarien unterscheiden. Buck und Limburg (2024) unterscheiden vier prototypische Nutzungsszenarien: KI-Tools als (a) *Ersatz* (d. h. KI als Denkersatz), (b) *Entlastung* (d. h. KI als Werkzeug), (c) *Unterstützung und emotionale Entlastung* (d. h. KI als Schreibbegleitung) und (d) *Erweiterung* (d. h. KI als Sparringpartnerin) menschlichen Denkens. Die Spannbreite reicht dabei, wie in Abbildung 22 dargestellt, von einem vollständigen Ersatz des menschlichen Denkens (passive Nutzung) bis zu einer qualitativen Erweiterung des menschlichen Denkens (aktive Nutzung). Die Nutzungsszenarien unterscheiden sich überdies u. a. in der Komplexität des Arbeitsprozesses, der Qualität der Ergebnisse und der Möglichkeit der Verantwortungsübernahme für den Text.

Abbildung 22: Prototypische Nutzungsszenarien von KI-Tools



Notizen. Die Abbildung illustriert die vier prototypischen Nutzungsszenarien von KI-Tools nach Buck und Limburg (2024). Die Nutzungsszenarien unterscheiden sich im Umfang des menschlichen Denkens. Beim Szenario *KI als Denkersatz* wird das menschliche Denken komplett an KI-Tools ausgelagert. Wird die *KI als Werkzeug* verwendet, bedeutet dies eine Abgabe von kognitiv weniger anspruchsvollen Teilaufgaben zur Entlastung des menschlichen Denkens. Beim Szenario *KI als Schreibbegleitung* wird das menschliche Denken durch KI-Tools durch die Lösung von während dem Schreiben auftretenden Problemen unterstützt. Das Szenario *KI als Sparringpartnerin* schlussendlich setzt die höchsten Ansprüche an das menschliche Denken – hier soll das Denken durch KI-Tools nicht entlastet oder unterstützt, sondern erweitert werden. Die Szenarien erfordern von links nach rechts nicht nur mehr menschliches Denken, sondern auch ausgeprägtere Schreibkompetenz und *AI literacy.*

1. KI als Denkersatz

- Szenario: Auslagerung des eigenen Denkens an KI-Tools
- Beispiel: Eine Aufgabenstellung eins zu eins als Prompt in ein KI-Tool hochladen und die Ausgabe ohne kritische Prüfung übernehmen (Buck & Limburg, 2024)
- Risiken: Die vollständig passive Nutzung von KI-Tools ist nicht zielführend. Um negative Auswirkungen zu verhindern, sollte bei der Nutzung von KI-Tools immer auch die eigene Denkarbeit, Expertise und Perspektive mit einfliessen (Dowling & Lucey, 2023; Khalifa & Albadawy, 2024; Salvagno et al., 2023). Andernfalls handelt es sich nicht um Schreiben im eigentlichen Sinn und es entsteht auch kaum ein Verantwortungsgefühl für den Text, da dieser (fast) komplett ohne eigenes Zutun geschaffen wurde. Auch ein Lernen im Sinne eines Kompetenzerwerbs wird durch eine solche Nutzung von KI-Tools vollständig verhindert (Buck & Limburg, 2024).

2. KI als Werkzeug

- Szenario: Abgabe kognitiv weniger anspruchsvoller Teilaufgaben an KI-Tools zur Entlastung des Denkens
- Beispiele: (a) Formulierung von Fliesstext aus Stichworten oder Textbausteinen, (b) Korrektur oder Umformulierung von Entwürfen, (c) Anpassung des Textstils (Buck & Limburg, 2024)
- Chancen: Die Entlastung des Denkens führt zu einer Effizienzsteigerung und schafft mehr Kapazitäten für kognitiv anspruchsvollere Teilaufgaben. Im Gegensatz zur KI als Denkersatz werden hierbei Teilaufgaben identifiziert, welche von KI-Tools ausgeführt werden können, und die Ausgabe wird hinsichtlich der Qualität überprüft.
- Anforderungen an Schreibende: Grundlegende Schreibkompetenz (Buck & Limburg, 2024)

3. KI als Schreibbegleitung

- Szenario: Lösung von konkreten, während dem Schreiben auftretenden Problemen mithilfe von KI-Tools
- Beispiele: (a) Unterstützung bei der Gliederung eines Textes oder beim Einstieg in einen Text, (b) Erklärung von Fachbegriffen oder Textstellen, (c) Finden von konkreten, geeigneten Formulierungen oder Synonymen (Buck & Limburg, 2024)
- Chancen: KI-Tools können Denkprozesse unterstützen und Schreibende emotional entlasten.
- Anforderungen an Schreibende: Grundlegende Schreibkompetenz sowie Wissen über die Möglichkeiten und Grenzen der genutzten KI-Tools (basale *AI literacy;* Buck & Limburg, 2024)

4. KI als Sparringpartnerin

- Szenario: Steigerung der inhaltlichen Qualität mithilfe von KI-Tools
- Beispiele: (a) Identifikation von Argumentationslücken, (b) Unterstützung beim Vergleich von Textversionen, (c) Einbringen von weiterführenden Ideen und Quellen (Buck & Limburg, 2024)
- Chancen: Es geht nicht um eine kognitive Entlastung, sondern die Nutzung von KI-Tools soll zu neuen Gedanken, zur Identifikation von Zu-

sammenhängen und zu einem tieferen Verständnis führen. Als Folge wird auch die Verantwortungsübernahme für den Text gestärkt (Buck & Limburg, 2024). Dieses Nutzungsszenario entspricht der bereits oben genannten hybriden Intelligenz, also der optimalen Kombination von Mensch und Maschine, wobei die Risiken von KI möglichst minimiert und die Chancen von KI maximiert werden (Rafner et al., 2021).

- Anforderungen an Schreibende: Um eine höhere inhaltliche Qualität zu erreichen, ist eine andauernde Überprüfung des eigenen Erkenntnisfortschritts notwendig. Dies erfordert eine *ausgeprägte Schreibkompetenz* sowie *AI literacy*. Schreibende sollten zudem nicht erwarten, mithilfe von KI-Tools in kürzester Zeit zu einem besseren Ergebnis zu gelangen. Die korrekte Nutzung von KI-Tools zum Schreiben kann viel Zeit in Anspruch nehmen, z. B. durch die Überprüfung von Aussagen oder die Steuerung des Schreibprozesses. Das Ziel sollte es sein, mit demselben zeitlichen Aufwand und ohne Einbussen beim Kompetenzerwerb ein besseres Ergebnis zu erzielen (Buck & Limburg, 2024). Zudem darf die Nutzung von KI als Sparringpartnerin nicht dazu verleiten, auf menschlichen Austausch – z. B. in Form von Anregungen oder Feedback durch Co-Autor:innen, Betreuungspersonen, Mitstudierende, Verwandte oder Freunde – zu verzichten. KI-Tools sollten immer als Ergänzung und nicht als Ersatz von menschlichem Austausch verwendet werden (Buck & Wessels, 2025).

Prompting

> "A well-crafted prompt can lead to a precise, accurate and relevant response, and thereby maximizes the model's performance. Conversely, a poorly structured prompt may result in a vague, incorrect or irrelevant answer."
>
> (Ein gut formulierter Prompt kann zu einer präzisen, zutreffenden und relevanten Antwort führen und damit die Leistung des Modells maximieren. Umgekehrt kann ein schlecht strukturierter Prompt zu einer vagen, falschen oder irrelevanten Antwort führen.)
>
> —Z. Lin (2024a, p. 611)

Ein Prompt ist eine sprachliche Instruktion an ein KI-Tool, mit welcher dieses zur Erledigung einer bestimmten Aufgabe bzw. Generierung einer bestimmten Ausgabe aufgefordert wird (Buck, 2025; Giray, 2023; Wallwork,

2024). Die Eingabe eines Prompts in ein KI-Tool wird dabei als Prompting bezeichnet.[17] Je nach Komplexität der Aufgabe reicht ein kurzer, einfacher Prompt in der Form einer Frage oder Anweisung aus, um das gewünschte Ergebnis zu erzielen. Insbesondere bei komplexen Aufgaben lohnt sich jedoch eine strukturiertere Herangehensweise.

Ein besonders strukturiertes Vorgehen beim Prompting wird auch als *prompt engineering* bezeichnet (Sahoo et al., 2025).[18] Es ist umstritten, wie nützlich *prompt engineering* tatsächlich ist, da KI-Tools sich zu einem gewissen Grad unvorhersehbar verhalten und somit mit *prompt engineering* erstellte Prompts nicht zwingend zu einer besseren Ausgabe führen (Mollick, 2024b).

> "The difficulty of prompt engineering disrupts the myth that GenAI programs can simply be told what to do with little thought on the part of the human user."
>
> (Die Schwierigkeit des *prompt engineering* widerlegt den Mythos, dass generativen KI-Programmen einfach gesagt werden kann, was sie tun sollen, ohne dass menschliche Nutzende darüber nachdenken müssen.)
>
> —Dobrin (2023, p. 74)

Durchaus hilfreich beim Prompting erweist sich jedoch die Nutzung von *ergänzenden Elementen,* welche situativ eingesetzt werden können (Mollick, 2024b). Um die Qualität der Ausgabe zu verbessern, werden dabei anstelle einer simplen Ausformulierung der Aufgabe auch bestimmte zusätzliche Anweisungen und Informationen in den Prompt inkludiert (Z. Lin, 2024a; Mollick, 2024b). Dazu gehören z. B. Angaben (a) zur Rolle, welche das KI-Tool übernehmen soll, (b) zum Kontext der Aufgabe sowie (c) zu formalen Aspekten der Ausgabe (vgl. Checkliste zum Prompting).

Ein solches Vorgehen bedingt Grundkenntnisse zum Prompting. Zwar werden KI-Tools besser dabei, auch auf ungenaue Prompts zu reagieren und zu «verstehen», was die Nutzenden tatsächlich beabsichtigen (Mollick, 2023b). Zudem werden KI-Tools durch die wachsende Grösse der zugrundeliegenden Sprachmodelle weniger sensitiv für Änderungen im Prompting (Zhuo et al., 2024). Dennoch lohnt es sich, einige Grundprinzipien des Promptings zu kennen.

---

17 Mittlerweile haben viele KI-Tools einen *voice mode,* mit welchem auch eine mündliche Eingabe (d. h. mündliches Prompting) möglich ist.

18 Bei vielen Begrifflichkeiten im Zusammenhang mit Prompting besteht jedoch noch kein Konsens zur genauen Definition (Schulhoff et al., 2025).

Grundprinzipien

> «Den *einen* perfekten Prompt gibt es nicht. Prompts sind keine Zaubersprüche, die man auswendig lernt und dann einfach in allen möglichen Situationen anwendet.»
>
> —Buck (2025, p. 100)

Anders als dies z. B. in gewissen Social Media-Posts oder Online-Artikeln impliziert wird, gibt es keine «magischen Prompts», welche immer funktionieren (Goedecke, 2025; Hirsch, 2023). Um mittels Prompting zur gewünschten Ausgabe zu gelangen, ist häufig ein sich wiederholender – d. h. mehrere sog. *Iterationen* umfassender – Vorgang notwendig. Auch detaillierte und hochgradig strukturierte Prompts führen nicht immer direkt zum Ziel. Die Ausgabe des KI-Tools wird daher im Verlauf der Konversation mit dem KI-Tool schrittweise verfeinert, bis das gewünschte Ergebnis erreicht ist (Buck, 2025; Ingley & Pack, 2023).

Als Startpunkt lassen sich *conversational prompting* und *structured prompting* unterscheiden (Mollick, 2023c).

- *Conversational prompting:* Das Ziel ist weniger klar definiert und daher der initiale Prompt weniger spezifisch formuliert. Die Konversation mit dem KI-Tool wird gezielt genutzt, um sich schrittweise einem zufriedenstellenden Ergebnis anzunähern (Mollick, 2023c). Dieses Vorgehen eignet sich eher für Aufgaben mit geringer Komplexität (Buck, 2025).
- *Structured prompting:* Das Ziel ist meist klar definiert und der initiale Prompt spezifischer und umfangreicher. *Structured prompting* bedingt mehr Vorwissen und Übung sowie eine genauere Auseinandersetzung mit dem gewünschten Ergebnis, ermöglicht aber auch bei anspruchsvollen Aufgaben eine hohe Qualität der Ausgabe (Buck, 2025; Mollick, 2023c).

Entspricht die erste Ausgabe des KI-Tools nicht dem gewünschten Ergebnis, können verschiedene Strategien angewendet werden. Welche Strategien zum Erfolg führen, lässt sich nur durch ein Experimentieren herausfinden (Dobrin, 2023; Z. Lin, 2024a). Manchmal führen bereits geringfügige Anpassungen in Prompts zu grossen Unterschieden in der Ausgabe (Z. Lin, 2024a). Auch die Wahl des KI-Tools spielt eine Rolle, da sich KI-Tools im Umgang mit Prompts voneinander unterscheiden – ein Prompt, welcher beim einen KI-Tool zum gewünschten Ergebnis führt, tut dies nicht zwingend bei einem anderen KI-Tool (Errica et al., 2025). Somit gibt es nicht «die eine» Lösung beim Prompting. Prompting braucht entsprechend Zeit

und eine gewisse Hartnäckigkeit (Mollick, 2023b). Schlussendlich hilft v. a. Übung dabei, die eigene Prompting-Kompetenz zu verbessern und den eigenen, passenden Prompting-Stil zu finden (Mollick, 2023b, 2024b).[19]

Die aufgelisteten Strategien können hilfreich sein, falls der initiale Prompt nicht zum gewünschten Ergebnis geführt hat.

Wiederholung des initialen Prompts

- Verwenden Sie den initialen Prompt erneut, ohne ihn zu verändern (Z. Lin, 2024a).
- Fordern Sie das KI-Tool auf, den initialen Prompt nochmals zu lesen (X. Xu et al., 2024).

Anpassung des initialen Prompts

- Sprachliche Anpassung: Strukturieren Sie den initialen Prompt um oder tauschen, ergänzen und/oder löschen Sie gewisse Wörter (Dobrin, 2023; Z. Lin, 2024a).
- Inhaltliche Anpassung: Passen Sie Instruktionen und Informationen an bzw. ergänzen und/oder löschen Sie Teile davon, z. B. Details zur Aufgabe, Restriktionen oder Beispiele (Dobrin, 2023; Z. Lin, 2024a).

Formulierung eines neuen Prompts

- Stellen Sie Fragen zur Ausgabe, nehmen Sie auf Aussagen des KI-Tools Bezug oder Antworten Sie auf Fragen des KI-Tools (Wallwork, 2024).
- Geben Sie konkretes Feedback dazu, wie die Ausgabe angepasst und verbessert werden soll (Mollick, 2023b).

Weitere Strategien

- Geben Sie den initialen Prompt in einer anderen Sprache ein, z. B. in Englisch, da in dieser Sprache in der Regel mehr Trainingsmaterial vorliegt (Buck, 2025)
- Geben Sie den initialen Prompt in ein anderes KI-Tool ein (Z. Lin, 2024a).

Darüber hinaus gibt es auch Strategien, welche nicht direkt an eine erste Ausgabe anknüpfen. Beispielsweise werden beim *chain of thought prompting*

19 Bei neueren, häufig zahlungspflichtigen Versionen lassen sich zudem z. B. Präferenzen hinterlegen und das KI-Tool «erlernt» aufgrund vergangener Konversationen den Prompting-Stil von Nutzenden (Varanasi, 2025).

(Meincke et al., 2024; Wei et al., 2023) komplexe Aufgaben bereits bei der Vorbereitung des Prompts in Unteraufgaben aufgeteilt und Schritt für Schritt abgearbeitet. Durch dieses schrittweise Vorgehen können Fehlerquellen entdeckt und korrigiert werden (Buck, 2025).

Checkliste zum Prompting

Die nachfolgende Checkliste bietet eine Übersicht zum Vorgehen beim Prompting (Buck, 2025; Giray, 2023; Ingley & Pack, 2023; Krüger, 2023; Z. Lin, 2024a; Mollick, 2023c, 2024b; Schulhoff et al., 2025; Wallwork, 2024). Die Checkliste eignet sich sowohl für *conversational prompting* als auch für *structured prompting*. Beim *conversational prompting* reicht aufgrund der geringeren Strukturierung eine oberflächliche Prüfung entlang der einzelnen Punkte der Checkliste. Beim *structured prompting* empfiehlt es sich, alle aufgelisteten Punkte genau zu prüfen.

☐ Vorbereitende Gedanken

> "AI is a tool. It is not always the right tool. Consider carefully whether, given its weaknesses, it is right for the purpose to which you are planning to apply it."
>
> (KI ist ein Werkzeug. Sie ist nicht immer das richtige Werkzeug. Prüfen Sie sorgfältig, ob KI angesichts ihrer Schwächen für den Zweck geeignet ist, für welchen Sie sie einsetzen wollen.)
>
> —Mollick (2023a, Section And More?)

- Art der Aufgabe:
  - Für welche Aufgabe möchte ich ein KI-Tool nutzen?
  - Beispiele: Beantwortung einer Frage, Erklärung eines Themas, Überarbeitung eines Textes
- Eignung für Aufgabe:
  - Sind KI-Tools für diese Aufgabe geeignet? Ist das gewählte KI-Tool für diese Aufgabe geeignet?
  - Erläuterung: KI-Tools sind nicht für alle Aufgaben geeignet und nicht jedes KI-Tool eignet sich für jede Aufgabe gleich gut
- Art des Promptings:
  - Welche Art des Promptings ist für diese Aufgabe am besten geeignet?

- Beispiele: *conversational prompting, structured prompting, chain of thought prompting*
- Kriterien zur Überprüfung:
  - Welche Kriterien sind zur Überprüfung der Ausgabe am besten geeignet?
  - Beispiele: Passung zur Aufgabenstellung, inhaltliche Konsistenz, Einbezug des Kontextes, Einhaltung formaler Vorgaben

☐ **Elemente des Prompts**

- Instruktion:
  - Ist die Instruktion prägnant und verständlich formuliert?
    - Gibt es z. B. noch schwer verständliche bzw. ambigue Begriffe? Ist die Instruktion spezifisch genug, um die Aufgabe ausführen zu können? Ist die Instruktion klar strukturiert?
- Rolle *(Persona):*
  - Soll das KI-Tool eine bestimmte Rolle einnehmen? Wenn ja, ist diese Rolle klar definiert?
  - Beispiele: Lehrperson, naive Leser:in, Reviewer:in, Lektor:in, *devil's advocate*
- Aufgabe:
  - Ist die Aufgabe klar definiert?
    - Wird bei komplexeren Aufgaben eine Schritt-für-Schritt Anleitung gegeben?
  - Beispiele: Beantwortung einer Frage, Erklärung eines Themas, Identifikation von Optimierungsmöglichkeiten, Überprüfung von Grammatik oder Stil
- Ggf. Kontext:
  - Gibt es Angaben zum Kontext, welche für die Aufgabe relevant sind? Wenn ja, welche?
    - Um welches Thema geht es? Welche Informationen oder Aspekte sollen in der Ausgabe unbedingt vorkommen? Was ist die Zielgruppe des Textes? Was soll mit der Ausgabe gemacht werden (d. h. Sinn und Zweck der Aufgabe)?
  - Beispiele: Abschlussarbeit an einer Hochschule, Bericht für Kund:innen aus der Automobilbranche, Zeitungsartikel zum Thema Gleichberechtigung
- Ggf. formale Aspekte (Restriktionen):
  - Gibt es formale Vorgaben für die Ausgabe?
    - Wie soll die *Tonalität* der Ausgabe sein?

        - Beispiele: Informell, formell, akademisch
    - In welcher *Textform* soll die Ausgabe sein?
        - Beispiele: Werbetext, Blogartikel, Tutorial, Geschichte, Abstract, Gedicht
    - In welchem *Format* soll die Ausgabe sein?
        - Beispiele: Fliesstext, Auflistung, Tabelle, PDF
    - Welchen *Umfang* soll die Ausgabe haben?
        - Beispiele: Fünf *bullet points,* 10 Sätze, 250 Worte
    - Soll der eingegebene Text *direkt korrigiert* oder *nur Feedback* ausgegeben werden?
        - Welche Überschriften sollen ggf. in der Ausgabe vorkommen?

☐ **Mögliche zusätzliche Inputs**

- Beispiele/Vorlagen:
    - Werden – falls möglich und sinnvoll – ein *(one-shot prompting)* oder mehrere *(few-shot prompting)* Beispiele/Vorlagen für die gewünschte Ausgabe mitgegeben?
- Text:
    - Wird der Text, mit welchem gearbeitet werden soll, mitgegeben?
- Weitere Informationen:
    - Werden weitere Informationen mitgegeben, welche das KI-Tool miteinbeziehen soll?
        - Wichtig ist hierbei, dass die Informationen möglichst genau bzw. korrekt sind. Zudem ist insbesondere auf das Urheberrecht und den Datenschutz zu achten (vgl. Ethische und rechtliche Rahmenbedingungen).
    - Beispiele: Eigene Daten, Dokumente, Expertise, Verweis auf frühere Konversationen

☐ **Nach Erhalt der Ausgabe**

- Überprüfung:
    - Entspricht die Ausgabe den Erwartungen?
        - Werden die zu Beginn definierten Kriterien erfüllt? Werden weitere Informationen benötigt? Muss die Ausgabe überarbeitet werden?
- Weiteres Vorgehen:
    - Welches Feedback kann dem KI-Tool gegeben werden, um die Ausgabe zu überarbeiten?

- Beispiele: Spezifizierung oder Hinzunahme von Elementen, Ergänzung um zusätzliche Inputs, Kombination verschiedener Aspekte von einer oder mehreren Ausgaben

Weiterführende Tipps

Prompting bedeutet auch, mit verschiedenen Strategien zu experimentieren. Es folgt eine nicht abschliessende Liste von weiterführenden Tipps für das Prompting. Die Liste soll als Inspiration dienen und umfasst Strategien, welche sich in bestimmten Situationen als wirksam erwiesen. Da Prompting oftmals zu Ausgaben führt, welche zumindest teilweise nützlich sind, ist die tatsächliche Wirksamkeit dieser Strategien – d. h. die inkrementelle Verbesserung ggü. anderen Strategien – jedoch schwer zu beurteilen. Umfangreiche Übersichten zu verschiedenen weiteren Prompting-Strategien und -Techniken finden sich z. B. bei Schulhoff et al. (2025) und Sahoo et al. (2025).

Erstellung/Überarbeitung von Prompts

- Verwenden Sie Prompt-Vorlagen bzw. bereits von anderen Personen genutzte Prompts als Startpunkt für Ihre eigenen Prompts (Schulhoff et al., 2025).
- Bitten Sie das KI-Tool, einen Prompt – in einer ersten Fassung – zu erstellen (Mollick, 2024b) oder Optimierungsmöglichkeiten zu einem bestehenden Prompt zu identifizieren (*meta-prompting;* Schulhoff et al., 2025; Wallwork, 2024).
- Fordern Sie das KI-Tool dazu auf, weiterführende Fragen zur Aufgabe zu stellen, bevor es eine Antwort ausgibt (Gibbs, 2025).
  - Beispiel: Welche weiteren Informationen wären für Dich hilfreich, damit du ein möglichst passendes Feedback zu meinem Text erstellen kannst?
- Verwenden Sie emotionale Appelle (Li et al., 2023).
  - Beispiel: Dieses Schreibprojekt ist sehr wichtig für meine Karriere, kannst du mich dabei unterstützen?
- Formulieren Sie (zumindest moderat) höfliche Prompts (Yin et al., 2024; siehe aber auch Meincke et al., 2025).
  - Beispiel: Bitte nenne mir die Hauptelemente von *AI literacy* und beschreibe diese kurz. Beschränke dich dabei bitte auf maximal drei Sätze pro Hauptelement.

- Fordern Sie das KI-Tool dazu auf, kreativ zu sein oder Annahmen zu treffen. Dies führt in vielen Fällen zu breiteren, neuartigeren Ausgaben (Mollick, 2023b).
  - Beispiel: Bitte triff alle Annahmen, welche du zur Beantwortung meiner Anfrage benötigst.
- Weisen Sie das KI-Tool darauf hin, dass es mitteilen soll, wenn es zu wenige Informationen hat oder «unsicher» ist. Dies hilft möglicherweise dabei, Fehler wie Halluzinationen in der Ausgabe verringern (Mollick, 2024b).
  - Beispiel: Bitte teile mit, wenn für die Beantwortung der Frage zu wenige Informationen verfügbar sind.
- Weisen Sie bei Textüberarbeitungen klar darauf hin, ob der Text direkt angepasst werden soll oder ob nur Vorschläge aufgelistet werden sollen (Wallwork, 2024).
  - Beispiel: Bitte passe den Text nicht direkt an, sondern liste alle deine Anpassungsvorschläge mit *bullet points* auf.
- Formulieren Sie eine Frage als Herausforderung oder versprechen Sie dem KI-Tool eine Belohnung, z. B. ein Trinkgeld (Bsharat et al., 2024).
  - Beispiel: Ich gebe dir ein Trinkgeld von 20$ für eine besonders kreative Lösung.

Verarbeitung der Ausgabe

- Fragen Sie das KI-Tool in einem ersten Schritt direkt nach Quellen, um die Qualität einer Ausgabe zu überprüfen und die Quellen mit der Ausgabe abzugleichen (Mollick, 2023b).
  - Beispiel: Bitte erstelle mir eine Liste aller für Deine Aussagen relevanten Quellen, inkl. Links.
- Bitten Sie das KI-Tool, die eigene Ausgabe zu bewerten und/oder Feedback dazu zu geben (*self-criticism;* Schulhoff et al., 2025).
  - Beispiel: Bitte gib ein Feedback zu deiner letzten Ausgabe und lege dabei den Fokus auf die Nachvollziehbarkeit der Argumentation.
- Wenn die Ausgabe – insbesondere im Hinblick auf die Genauigkeit – nicht den Erwartungen entspricht, geben Sie einen Prompt wiederholt ein und lassen Sie die Ausgaben vom KI-Tool zu einer «Gesamt-Ausgabe» aggregieren (*ensembling;* Schulhoff et al., 2025).
  - Beispiel: Bitte fasse deine 10 Ausgaben zum Prompt «[Platzhalter Prompt]» zu einer Gesamt-Ausgabe zusammen.

Sonstige Massnahmen

- Variieren Sie die *temparature*-Einstellungen. Je nach Art der Aufgabe führt eine Erhöhung (höhere Kreativität) oder Verringerung (höhere Genauigkeit) der *temperature* zu einer besseren Ausgabe (Buck, 2025).
- Öffnen Sie ein neues Chatfenster, wenn sich die Interaktion mit dem KI-Tool nicht wie gewünscht entwickelt – d. h. wenn Sie das Gefühl haben, in eine «Sackgasse» geraten zu sein (Laban et al., 2025; Mollick, 2023b).

Anwendungsbereiche entlang des Schreibprozesses

KI lässt sich auch entlang des Schreibprozesses unterschiedlich nutzen. Allgemein können zwei Anwendungsformen beim Schreiben mit KI-Tools unterschieden werden: Die Arbeit *mit dem Text* und die Arbeit *ohne Text*. Bei der Arbeit mit dem Text wird ein bestimmter Text (z. B. ein Paragraf, eine Gliederung) in ein KI-Tool hochgeladen. Das KI-Tool gibt dann eine Rückmeldung zu definierten Aspekten oder eine nach definierten Kriterien überarbeitete Version des Textes aus. Bei der Arbeit ohne Text wird kein Text hochgeladen. Stattdessen soll z. B. ein Vorschlag für Ideen oder für eine Gliederung ausgegeben werden. Dieser Vorschlag wird dann mit den eigenen Ideen oder der eigenen Gliederung verglichen, um diese zu erweitern oder zu optimieren.

Nachfolgend werden Beispiele für mögliche Anwendungsbereiche und Aufgaben entlang des Schreibprozesses aufgelistet, bei welchen KI-Tools Unterstützung bieten können. Die Auflistung basiert teilweise auf bestehenden Auflistungen (Buck, 2025; Buck & Limburg, 2024; Dobrin, 2023; Hutson, 2025; Ingley & Pack, 2023; Kosberg, 2024; Lingard, 2023; Wallwork, 2024) und orientiert sich an den Schritten eines Schreibprojekts (vgl. Kapitel 2.1 Phasenmodell des Schreibprozesses).

Die Generierung von Text durch KI-Tools wird bei der untenstehenden Auflistung bewusst ausgeklammert. Die direkte Textgenerierung ist bei Schreibprojekten nicht zielführend, weil dadurch ein essenzielles Element des Schreibens verloren geht *(writing is thinking on paper)*. Mögliche Folgen dieser Nutzung von KI als Denkersatz (siehe oben) wären eine geringere Textqualität, eine Gefährdung des Erwerbs und Erhalts von Schreibkompe-

tenz und eine fehlende oder geringere Verantwortungsübernahme für den Text.[20]

> "Always generate your own ideas before turning to AI. Write them down, no matter how rough. Just as group brainstorming works best when people think individually first, you need to capture your unique perspective before AI's suggestions can anchor you."
>
> (Entwickeln Sie immer Ihre eigenen Ideen, bevor Sie KI nutzen. Schreiben Sie die Ideen auf, egal wie unausgereift. Genauso wie ein Gruppen-Brainstorming am besten funktioniert, wenn die Teilnehmenden zuerst individuell denken, müssen Sie Ihre einzigartige Perspektive festhalten, bevor die Vorschläge der KI Sie ankern können.)
>
> —Mollick (2025, Section The Creative Brain)

Sobald Sie ein KI-Tool nach etwas gefragt und eine Antwort darauf erhalten haben, gibt es kein Zurück mehr. Durch diese Ankersetzung ist es Ihnen nicht mehr möglich, eigene Gedanken zu formulieren, welche von der Ausgabe des KI-Tools vollkommen unabhängig sind. Die Nutzung von KI-Tools beginnt somit immer mit den eigenen Gedanken und Formulierungen (Mollick, 2025). Erst in einem zweiten Schritt werden KI-Tools als Unterstützung hinzugezogen (Dobrin, 2023; Ingley & Pack, 2023; Oertner, 2024).

Richtlinie: Beginnen Sie den Prozess stets mit ihren eigenen Gedanken und halten Sie diese schriftlich oder ggf. als Audiodatei fest. "You never get a second chance to note your first thoughts."

<u>Fragestellung und Recherche</u>

- Ideenfindung und Brainstorming
- Sammlung erster Informationen zu einem Thema *(Einstiegssuche)*
- Identifikation (a) geeigneter Schlagworte und Datenbanken zur Durchführung einer systematische Literaturrecherche und (b) weiterführender Quellen zur breiten Erfassung eines Themas
  - <u>Hinweis:</u> Führen Sie stets eine eigene Recherche durch, um sich vertieft mit einem Thema zu befassen und nichts zu übersehen. Setzen Sie KI-Tools lediglich ergänzend zur Recherche ein.

20 Darüber hinaus gibt es z. B. im Hochschulkontext auch Vorgaben dazu, wie KI-Tools von Studierenden genutzt werden dürfen. Eine Generierung von Text mit KI-Tools ist dabei teilweise explizit untersagt.

- Beispielprompt: Du bist ein Experte zum Thema KI. Ich möchte ein Buchkapitel zum Thema «Schreiben mit KI» verfassen. Folgende Unterthemen möchte ich in das Kapitel integrieren: [Auflistung Unterthemen]. Gib mir bitte Rückmeldung dazu, welche weiteren Ideen ich in das Kapitel einbringen könnte und welche Aspekte bei den aufgelisteten Unterthemen vorkommen sollten. Liste dann alle relevanten Unterthemen inkl. relevanter Aspekte, die im Kapitel vorkommen, als *bullet points* auf.

Notizen

- Überführung gesprochener Sprache, Stichworte oder handschriftlicher Notizen in einen (digitalen) Fliesstext
- Unterstützung bei Verständnis und Abgrenzung von Konzepten und Fachbegriffen
- Weiterführende Auseinandersetzung mit Notizen zur Verdichtung und einem vertieften Verständnis, z. B. alternative Ansichten über ein Thema, Argumente und Gegenargumente für eine bestimmte Position
- Beispielprompt: Du bist eine Expertin zum Thema KI. Ich möchte ein Buchkapitel zum Thema «Schreiben mit KI» verfassen. Kannst du mir bitte eine für Laien verständliche Erklärung geben, wie sich unüberwachtes Lernen und verstärkendes Lernen voneinander unterscheiden? Ich interessiere mich v. a. für die Unterschiede in der Funktionsweise. Gib mir die Informationen als Fliesstext aus und beschränke dich auf maximal 200 Wörter.

Gliederung *(Outline)*

- Überprüfung einer selbst erstellten Gliederung auf (a) Vollständigkeit und (b) Konsistenz der Titelsetzung
- Vorschläge für Gliederungsvarianten und Titelsetzung
- Vorschläge zur Ordnung von Kategorien (Umbenennung, Zusammenführung, logische Reihenfolge)
- Anreicherung der Gliederung durch z. B. Beispiele, Metaphern, Evidenz oder Erläuterungen aus den Notizen
- Beispielprompt: Du bist ein Experte zum Thema «Schreiben mit KI». Ich möchte ein Buchkapitel zu diesem Thema verfassen. Dazu möchte ich in einem ersten Schritt eine Gliederung erstellen. In der Gliederung sollen folgende Punkte vorkommen: [Auflistung Gliederungspunkte]. Bitte erstelle mir drei Vorschläge zur Gliederung des Kapitels. Erläutere zudem jeweils, welches die Vor- und Nachteile der Gliederung sind. Gib alles in

Form einer Vergleichstabelle aus, in welcher jede Spalte einen der drei Vorschläge zur Gliederung des Kapitels beinhaltet.

Paragrafen

- Überprüfung der Organisation innerhalb eines Paragrafen (vgl. Checkliste Paragrafen aus Kapitel 2.3 Arbeit mit Entwürfen)
- Überprüfung der (a) Kohärenz und Stringenz und (b) der Klarheit und Verständlichkeit eines Paragrafen
- Überprüfung der (a) Reihenfolge von Paragrafen bzw. der Argumentationslinie *(train of thought)* und (b) der Argumentationslinie auf Lücken
- Beispielprompt: Du bist eine Lehrperson und möchtest dich zum Thema «Schreiben mit Künstlicher Intelligenz» informieren. Du verfügst über keinerlei Vorwissen zum Thema. Gib mir bitte Feedback dazu, wie klar und verständlich der nachfolgende Text für dich ist. Gib dabei insbesondere die Textstellen an, welche du nicht verstehst und erläutere, was du daran nicht verstehst. Hier folgt der Text: «[Platzhalter Text]»

Überarbeitung *(Revision)*

- Überprüfung der Argumentationslinie
- Vorschläge für mögliche inhaltliche Anpassungen und Ergänzungen (fehlende Informationen, Widersprüche, Redundanzen)
- Feedback (a) zu Optimierungsmöglichkeiten der Paragrafen (Entfernung, Hinzufügung oder Kürzung von Paragrafen) und (b) zur Platzierung der Paragrafen in der Argumentationslinie
- Überarbeitung von (a) Formulierungen zur Verbesserung der Verständlichkeit und (b) Überschriften (Platzierung, Formulierung, Passung zum Abschnitt)
- Beispielprompt: Du bist ein Experte zum Thema «Schreiben mit Künstlicher Intelligenz». Ich möchte ein Buchkapitel zu diesem Thema schreiben. Gib mir bitte Feedback dazu, ob im nachfolgenden Text im Abschnitt «[Name Abschnitt]» wichtige Informationen fehlen und ob Widersprüche oder Redundanzen vorhanden sind. Liste mir dazu alle fehlenden Informationen, Widersprüche und Redundanzen auf und begründe deine Einschätzung jeweils. Hier folgt der Text: «[Platzhalter Text]»

Lektorat *(Copyediting)* und Korrekturlesen *(Proofreading)*

- Überprüfung von (a) Sprachstil, (b) adäquater und konsistenter Nutzung von Fachbegriffen, (c) Abkürzungen und (d) allg. Verständlichkeit
- Vorschläge für (a) Aufteilung von langen und komplexen Sätzen in kürzere Sätze, (b) Übergänge zwischen Sätzen oder Paragrafen und (c) (verständliche) Synonyme
- Überprüfung von (a) Grammatik (inkl. Zeichensetzung) und (b) Rechtschreibung
- Beispielprompt: Du bist ein erfahrener Lektor bei einem grossen Buchverlag. Überprüfe bitte im folgenden Textabschnitt, ob Begriffe konsequent gleich verwendet werden. Liste mir dazu alle Textstellen auf, in welchen ein Begriff unpassend verwendet wird und mache mir einen Vorschlag, wie der Text umgeschrieben werden könnte. Hier folgt der Text: «[Platzhalter Text]»

Übergeordnete Aufgaben

- Schreibprojektmanagement:
  - Identifikation von Forschungsfrage, Ziel und Zweck
  - Klärung von Anforderungen, Erwartungen und Bedürfnissen
  - Identifikation von relevanten Teilaufgaben und Meilensteinen
- Umgang mit Schreibblockaden über bereits bei spezifischen Phasen genannter Unterstützung hinaus, anknüpfend an Kapitel 2.4 Umgang mit Schreibblockaden
  - Unterstützung beim *free writing,* d. h. produzierten Text in Fliesstext umwandeln
  - Schreibmotivation durch emotionale Unterstützung
  - Gespräch mit KI-Tool als Erweiterung des Selbstgesprächs/Selbstinterviews, inkl. ggf. Aufnahme und Transkription durch das KI-Tool

7.4 Ausblick

Inhaltsübersicht

KI wird bleiben

> "The spirit of innovation is reckless, seeking constant change and flux, creating uncertainty in its wake."
> (Der Geist der Innovation ist rücksichtslos, er strebt nach ständiger Veränderung und erzeugt Ungewissheit in seinem Kielwasser.)
>
> —Saha (1998, pp. 511–512)

Auch wer sich in die Vergangenheit flüchtet, wird die Zukunft nicht aufhalten können. – Die Präsenz von KI ist die neue Realität, an welche sich die Gesellschaft gewöhnen muss, ähnlich wie sie sich bereits an frühere bahnbrechende Innovationen wie z. B. den Buchdruck, das Auto oder das Internet gewöhnen musste. KI hat bereits jetzt und wird auch weiterhin Auswirkungen auf alle Lebensbereiche haben. Es gilt, KI möglichst gewinnbringend einzusetzen und gleichzeitig den Risiken vorzubeugen, welche KI zweifellos mit sich bringt.

Wo der Weg mit KI genau hinführen wird, ist unklar.[21] KI-Tools werden weiter rasant verbessert, Erwartungen übertroffen. Ein Ende des KI-Hypes ist nicht in Sicht.[22]

21 Eine Übersicht über mögliche Zukunftsszenarien des Schreibens in der Wissenschaft und deren Auswirkungen auf die Gesellschaft – in Form von Utopien und Dystopien – bieten Limburg et al. (2023).

22 Es sei darauf hingewiesen, dass ein Teil dieses Hypes auch durch geschicktes Marketing der Anbieter von KI-Tools zustande kommt (Filipović et al., 2025).

> «Die Schwäche des Trainingsmaterials wird in Zukunft vermutlich weiter zunehmen. Grund dafür ist der unstillbare Hunger der neuen Grossen *[sic]* Sprachmodelle (LLM). Der Textbedarf ist jetzt bereits grösser als der Datenzuwachs im Internet, weshalb er durch KI-generiertes Material ergänzt werden muss. Hinzukommt, dass auch der Internet-Content zunehmend selbst KI-generiert ist … . Die KI füttern sich also direkt und indirekt gegenseitig, was die Verzerrungen potenziert: 'Garbage in, out, in, out, in, and out again.'»
>
> —Oertner (2024, p. 282)

KI-Tools können langfristig nicht immer noch besser werden, zumindest nicht im gleichen Tempo wie bislang. Zwar ist es möglich, gewisse Fehler zu beheben, z. B. durch Filter oder bessere zugrundeliegende Sprachmodelle (Oertner, 2024).[23] Es wird aber immer Fehler geben. Dies liegt an der Funktionsweise von KI-Tools, insbesondere am genutzten Trainingsmaterial und dem Zufallsfaktor bei der Textgenerierung (Z. Xu et al., 2025). Die Qualität des Trainingsmaterials lässt sich zwar durch bessere Selektion und Überprüfung erhöhen, was zu einer Verbesserung der Funktionsweise von KI-Tools führen würde. Aber für grössere Sprachmodelle ist immer mehr von Menschen erstelltes[24] Textmaterial notwendig, welches gar nicht in dieser Zahl in ausreichender Qualität vorhanden ist (Oertner, 2024).

> "As AI keeps improving, the number of problematic cases keeps shrinking and thus it becomes more usable. However, the problematic cases never seem to disappear. It's as if you took a step that brings you 80% of the way toward a destination, and then another step covering 80% of the remaining distance, and then another step to get 80% closer, and so on; you'd keep getting closer to your destination but never reach it."
> (Mit der Verbesserung der künstlichen Intelligenz sinkt die Zahl der problematischen Fälle, so dass sie immer besser nutzbar wird. Allerdings scheinen die problematischen Fälle nie zu verschwinden. Es ist so, als ob man einen Schritt macht, der einen 80 % des Weges zu einem Ziel bringt, und dann einen weiteren Schritt, der 80 % der verbleibenden Strecke zurücklegt, und dann einen weiteren Schritt, um 80 % näher zu kommen, und so weiter; man kommt dem Ziel immer näher, erreicht es aber nie.)
>
> —Maggiori (2023, Chapter 3)

23 Einige der bei den Risiken von KI berichteten Befunden werden in zukünftigen Versionen der KI-Tools kaum mehr auftreten oder treten – aufgrund des rasanten Tempos der Weiterentwicklung – bereits zum jetzigen Zeitpunkt kaum mehr auf (vgl. Filipović et al., 2025).

24 Wird KI-generiertes Textmaterial beim Training eines KI-Tools verwendet, kann dies zu einer Verschlechterung der Leistung führen (Shumailov et al., 2024).

Neuere Versionen von KI-Tools sind zudem nicht in jedem Fall besser als ältere Versionen. Im April 2025 musste z. B. OpenAI ein Update zurückziehen, weil die neue Version den Nutzenden zu sehr «gefallen wollte» *(sycophancy)* und dies zu unterschiedlichen Problemen führte (z. B. Verstärkung von Verschwörungstheorien und gesundheitsgefährdendem Verhalten; Klein, 2025; OpenAI, 2025c). Updates können darüber hinaus auch zu neuen Fehlern führen, was im Phänomen des *catastrophic forgetting* («katastrophales Vergessen») begründet ist: Bei der Überarbeitung von KI-Tools – d. h. wenn die zugrundeliegenden Sprachmodelle Neues «lernen» – geht altes Wissen verloren (French, 1999). Dies scheint bei grösseren Sprachmodellen sogar verbreiteter zu sein als bei kleineren Sprachmodellen (Luo et al., 2025). Da die Bestrebungen momentan mehrheitlich in die Richtung gehen, Sprachmodelle immer noch grösser zu machen, ist damit zu rechnen, dass dieses Phänomen zukünftig noch häufiger vorkommen wird.

Die kritische Überprüfung der Ausgabe von KI-Tools wird damit weiterhin wichtig bleiben. Sie wird künftig sogar wichtiger, da Fehler wie Halluzinationen aufgrund der Verbesserungen bei den KI-Tools immer schwieriger zu entdecken sein werden – u. a. weil KI-Tools vermehrt auf sog. *reasoning models* basieren, welche noch überzeugender argumentieren können als bisherige Modelle (Z. Sun et al., 2025).

Massnahmen zum Umgang mit KI

Um mit der neuen Realität von KI umzugehen und damit einhergehenden Risiken vorzubeugen, sind verschiedene Massnahmen notwendig. Durch diese Massnahmen sollen (a) unerwünschte Kompetenzverluste verhindert, (b) Kompetenz im Umgang mit KI *(AI literacy)* aufgebaut und (c) geeignete Rahmenbedingungen für die Nutzung von KI geschaffen werden.

Kompetenzverluste verhindern

> «Das spezifische Risiko von Kompetenzverlusten im Zusammenhang mit dem Einsatz von KI-Anwendungen liegt demnach in der Besonderheit der Tätigkeiten, die der Technik überlassen werden – handelt es sich dabei um gesellschaftlich besonders bedeutsame oder kritische Einsatzbereiche, ist ein Verlust von menschlichen Kompetenzen und Fertigkeiten ein ernstzunehmendes Risiko.»
>
> —Deutscher Ethikrat (2023, p. 355)

Insbesondere auf gesellschaftlicher Ebene bedeutet der Vormarsch von KI, dass unerwünschte Kompetenzverluste verhindert werden müssen. Wie bei anderen technologischen Entwicklungen werden gewisse Kompetenzen mit der Zeit an Relevanz verlieren und allmählich verloren gehen, andere Kompetenzen hingegen wichtiger werden (Reinmann, 2023).

Diese Entwicklung ist nicht per se schlecht – solange sie bewusst geschieht. Beim Schreiben wurde vor ein paar Dekaden in der Schule noch sehr viel Wert auf das Schönschreiben gelegt. Heutzutage liegt der Fokus – auch durch die Digitalisierung – auf anderen Aspekten des Schreibens. Allerdings schien der Trend in gewissen Ländern wie z. B. Schweden zu weit gegangen zu sein. Zwischenzeitlich wurde dort in Schulen schwerpunktmässig mit Tablets gearbeitet, mittlerweile wird wieder vermehrt auf analoges Schreiben gesetzt – weil durch die einseitige Nutzung von digitalen Hilfsmitteln z. B. wichtige motorische Fähigkeiten verloren gehen können (Höfler, 2024).

> "The question now is not whether AI will change education, but how we will shape that change to create a more effective, equitable, and engaging learning environment for all."
>
> (Die Frage ist nicht, ob KI das Bildungswesen verändern wird, sondern wie wir diesen Wandel gestalten werden, um ein effektiveres, gerechteres und engagierteres Lernumfeld für alle zu schaffen.)
>
> —Mollick (2024a, Section Encouraging, Not Replacing, Thinking)

Das bedeutet, dass identifiziert werden muss, welche Kompetenzen erhalten bleiben und ggf. gestärkt werden und welche Kompetenzen neu aufgebaut werden sollen. Im Zusammenhang mit dem Schreiben kann postuliert werden, dass Schreibkompetenz weiterhin vermittelt werden sollte und sowohl Informationskompetenz – insbesondere zur Verifikation von Informationen – als auch kritisches Denken gestärkt werden sollten (Gerlich, 2025; H.-P. Lee et al., 2025; Reinmann, 2023).

Damit diese Kompetenzsicherung gelingt, sind auch regulatorische Massnahmen notwendig (Reinmann, 2023). Insbesondere Bildungsministerien sind dabei in der Pflicht, den Kompetenzerwerb auf Primar-, Sekundar- und Tertiärstufen sicherzustellen (vgl. Filipović et al., 2025).

> «Bereitstellung und Nutzung von Textgeneratoren gefährden im Extremfall ganze Berufe und Branchen. So ergeben sich Risiken für Journalisten, Autoren und Texter. Lange Zeit galten kreative und künstlerische Berufe als geschützt, was weder im Textbereich noch im Bildbereich mehr der Fall ist, zumindest dort nicht, wo Massenprodukte und Durchschnittsware ausreichen.»
>
> —Bendel (2024, p. 303)

Der Vormarsch von KI bedeutet auch, dass sich gewisse Arbeitsfelder bereits ändern oder zukünftig noch ändern werden. Im Zusammenhang mit dem Schreiben gehören dazu insbesondere klassische Bereiche wie z. B. Journalismus und Marketing (Bendel, 2024), aber auch z. B. gewisse juristische Arbeitsfelder (Schwarcz et al., 2025).

Auch in vielen anderen Arbeitsfeldern wird es zu einer Verschiebung von Aufgaben kommen (vgl. Ooi et al., 2025). Im Gesundheitsbereich z. B. könnten (zuverlässig funktionierende) KI-Tools weite Verbreitung bei niederschwelligen medizinischen Beratungen finden und so das Gesundheitssystem entlasten. Bereits jetzt scheinen KI-Tools gleich gut (oder sogar besser) bei der Beantwortung von per Chat gestellten medizinischen Fragen und bei medizinischen Diagnosen zu sein als medizinisches Fachpersonal (Arora et al., 2025; Everett et al., 2025). Dieses Fachpersonal könnte langfristig entlastet werden und somit «frei» werden für Aufgaben, welche nicht von KI-Tools erledigt werden können, wie z. B. medizinische Untersuchungen und Eingriffe sowie Pflege (Deutscher Ethikrat, 2023). Zu beachten ist hier allerdings, dass das medizinische Fachpersonal nicht vollständig «aus dem Loop» genommen werden darf und medizinische Laien nur noch mit KI-Tools kommunizieren. Das gesundheitliche Risiko wäre in solchen Fällen zu gross (Fang & Perkins, 2024).

AI literacy aufbauen

> "The concept of AI literacy emerges as a cornerstone of contemporary learning. In its essence, it deals with the understanding and capability to interact effectively with AI technology."
>
> (Das Konzept der KI-Kompetenz tritt als ein Eckpfeiler des modernen Lernens hervor. Im Kern geht es dabei um das Verständnis und die Fähigkeit, effektiv mit KI-Technologie zu interagieren.)
>
> —Walter (2024a, p. 11)

Über die Verhinderung von unerwünschten Kompetenzverlusten hinaus soll gleichzeitig neu eine Kompetenz vermittelt werden, welche im Umgang mit KI relevant ist. Zu dieser *AI literacy* gehören (a) ein grundsätzliches Verständnis der Funktionsweise, (b) Kenntnis der ethischen und rechtlichen Rahmenbedingungen sowie (c) der sich aus der Nutzung von KI ergebenden Chancen und Risiken und (d) die Fähigkeit, KI bei Aufgaben praktisch anzuwenden (Ng et al., 2021; Southworth et al., 2023).

AI literacy soll Nutzende befähigen, verantwortungsvoll mit KI umzugehen. Mit der Vermittlung dieser Grundkompetenz kann auch verhindert werden, dass falsche Vorstellungen von den Fähigkeiten von KI-Tools entstehen (Bender, 2024) oder die Nutzung von KI-Tools als überfordernd wahrgenommen wird. Ziel ist somit auch ein gesundes Misstrauen und eine Handlungsfähigkeit im Umgang mit KI-Tools.

Auch hier sind Bildungsministerien gefordert: Sie sollen diese Grundkompetenz fördern – einerseits durch direkte Vermittlung durch geschulte Lehrende auf Primar-, Sekundar- und Tertiärstufen, andererseits durch öffentliche Informationskampagnen (Filipović et al., 2025; Kultusministerkonferenz der Bundesrepublik Deutschland, 2025). Hochschulen und andere Forschungsinstitutionen können in diesem Zusammenhang auch durch eine Zusammenarbeit mit Medien in Form von *Wissenschaftskommunikation* einen Beitrag leisten, indem sie die Öffentlichkeit über relevante Forschungsergebnisse im Bereich KI informieren.

Rahmenbedingungen schaffen

Ein verantwortungsvoller Umgang mit KI bedeutet auch, dass geeignete Rahmenbedingungen geschaffen werden (Bengio et al., 2024; Deutscher Ethikrat, 2023). Relevant sind u. a. (a) Urheberrecht und Datenschutz, (b) Transparenz und Fairness sowie (c) Verhinderung von schädlicher Verwendung.[25]

25 Es wird hier v. a. auf Aspekte eingegangen, welche spezifisch für das Schreiben mit KI relevant sind. Allgemeinere Herausforderungen von KI werden dabei ausgeklammert. Dazu gehören z. B. ethische und arbeitsrechtliche Probleme bei der Datenaufbereitung (u. a. Ausbeutung von Arbeitskräften, unmenschliche Arbeitsbedingungen), die Überwachung durch Staaten oder private Organisationen, die Förderung von Chancenungleichheit sowie schädliche Einflüsse auf die Umwelt (Bender, 2024; Crawford, 2024; Luccioni et al., 2025).

Gewisse politische Bestrebungen wurden bereits umgesetzt, um diese Aspekte zu adressieren. Beispielsweise hat der Europarat 2024 eine KI-Konvention verabschiedet und die EU 2025 den *Artificial Intelligence Act* in Kraft gesetzt, welcher die Verbreitung und Nutzung von KI regulieren soll. Da es sich bei den Anbietern von KI-Tools häufig um private, global tätige Unternehmen handelt, ist eine politische Regulierung jedoch lediglich in beschränktem Rahmen möglich. Zudem besteht das Risiko, dass zu strenge Regulierungen den Fortschritt in unnötiger Weise behindern. Daher sind zwar zumindest gewisse grundlegende Regulierungen notwendig, um den Risiken von KI vorzubeugen. Gleichzeitig bleibt aber offen, wie weitgehend diese Regulierungen sein sollen und inwieweit sich umgekehrt der Markt selbst regulieren soll (Ferretti, 2022; Walter, 2024b). Die Auseinandersetzung mit den passenden Rahmenbedingungen wird durch die Schnelllebigkeit im Bereich KI auf jeden Fall weiterhin ein Thema bleiben.

Urheberrecht und Datenschutz:

- Welche Trainingsdaten dürfen von KI-Tools unter welchen Bedingungen genutzt werden? Es stellt sich die Frage, ob auch urheberrechtlich geschützte Texte als Trainingsdaten verwendet werden dürfen. Je nach Auslegung ist dies gemäss geltendem Recht unter bestimmten Voraussetzungen erlaubt, die aktuelle rechtliche Handhabung ist jedoch z. B. in den USA und in der EU umstritten (Dornis & Stober, 2024; Rankin, 2025; United States Copyright Office, 2025b).
- Wie wird der Datenschutz gewährleistet? Zum einen sollen sensible Daten, welche aus unterschiedlichen Gründen in den Trainingsdaten vorkommen, geschützt werden. Zum anderen geht es auch um alle Daten, welche von Nutzenden in KI-Tools eingegeben werden.

Transparenz und Fairness:

- Wie kann mehr Transparenz zur Funktionsweise von KI-Tools erreicht werden? Durch unbekannte Trainingsdaten und Algorithmen ist die Funktionsweise von KI-Tools für Aussenstehende häufig nicht nachvollziehbar. Zwar existieren einige *open source* Sprachmodelle, bei welchen alle Parameter offengelegt werden (z. B. BLOOM von HuggingFace; Scao et al., 2023). Die «grossen» KI-Tools und deren zugrunde liegenden Sprachmodelle müssen jedoch allesamt in ihrer Funktionsweise als zumindest teilweise intransparent bezeichnet werden. Mehr Transparenz in

Form von *erklärbarer KI* (*explainable artificial intelligence;* Samek et al., 2021) lässt sich z. B. durch neue rechtliche Rahmenbedingungen oder durch die Förderung von KI-Tools mit transparenter Funktionsweise erreichen. Der *Artificial Intelligence Act* der EU ist ein Schritt in diese Richtung, da Anbieter von KI-Tools gewisse Auflagen bezüglich Transparenz erfüllen müssen.

- Wie kann Fairness in der Nutzung von KI-Tools sichergestellt werden? Hier geht es insbesondere darum, Verzerrungen in den Ausgaben von KI-Tools vorzubeugen oder zumindest zu verringern (Ferrara, 2023a). Dazu sind u. a. möglichst unverzerrte Trainingsdaten notwendig (Bender, 2024). Inwieweit Trainingsdaten durch rechtliche Rahmenbedingungen verbessert werden können, ist allerdings unklar. Ein möglicher Ansatz zur Verringerung von Verzerrungen wäre die gezielte Förderung von Forschung zu Verzerrungen bei KI-Tools (European Union Agency for Fundamental Rights, 2022).

Verhinderung von schädlicher Verwendung:

- Wie kann Cyberkriminalität mithilfe von KI-Tools verhindert werden? Im Bereich der Cyberkriminalität bestehen bereits gewisse rechtliche Rahmenbedingungen, welche jedoch im Zusammenhang mit KI-Tools allenfalls erweitert werden müssen.

  > «Durch die Zunahme von authentisch wirkenden Texten wird das Risiko von Des- und Misinformation verstärkt und dadurch auch Schutzgüter von öffentlichem Interesse, wie der Schutz der Demokratie oder der Rechtsstaatlichkeit, gefährdet.»
  >
  > —Jahnel und Heil (2024, p. 347)

- Wie kann die Verbreitung von *Mis- und Desinformationen* mithilfe von bzw. durch KI-Tools verhindert werden? Misinformation meint hier die Verbreitung von falschen Informationen ohne Täuschungsabsicht. Dieses Risiko besteht insbesondere durch falsche Ausgaben von KI-Tools (Halluzinationen). Desinformation bezeichnet die bewusste Verbreitung von falschen Informationen mit Täuschungsabsicht. Mithilfe von KI-Tools ist Desinformation deutlich einfacher möglich. Sowohl Mis- als auch Desinformation wird somit durch KI-Tools begünstigt. Daher müssen auch in diesem Zusammenhang mögliche Gegenmassnahmen diskutiert werden, auch wenn sich diese als schwierig umsetzbar erweisen dürften.

- Für welche weiteren Bereiche muss der Einsatz von KI-Tools reguliert werden? Hier geht es insbesondere um Bereiche, für welche KI-Tools aufgrund deren Funktionsweise weniger geeignet sind oder eine falsche Anwendung respektive durch KI ausgelöste Fehler grosse Risiken mit sich bringen. Ein Beispiel wäre der Gesundheitsbereich, bei welchem Diagnosen durch Laien ohne zusätzliches Einholen von Expertise gravierende gesundheitliche Folgen haben können (Fang & Perkins, 2024; für ein konkretes Beispiel, siehe Eichenberger et al., 2025).

Fazit

KI ist gekommen, um zu bleiben. Wohin die Reise mit KI führt, hängt davon ab, wie die Gesellschaft mit den mit KI einhergehenden Herausforderungen umgeht. Das Best-Case-Szenario wäre eine ideale Kombination von Mensch und Maschine in Form von *hybrider Intelligenz:* Mensch und KI ergänzen sich optimal und führen so zu einer besseren Welt. Das Worst-Case-Szenario wäre ein Verlust von gesellschaftlich zentralen Kompetenzen und damit einhergehend eine Gefährdung der Gesellschaft als Ganzes und der demokratischen Grundordnung. Ein verantwortungsvoller Umgang mit KI-Tools ist notwendig, damit KI sich auf individueller und gesellschaftlicher Ebene (überwiegend) positiv auswirken kann.

Literatur

Agrawal, A., Suzgun, M., Mackey, L., & Kalai, A. T. (2024). *Do language models know when they're hallucinating references?* (No. arXiv:2305.18248). arXiv. https://doi.org/10.48550/arXiv.2305.18248

Airenti, G. (2015). The cognitive bases of anthropomorphism: From relatedness to empathy. *International Journal of Social Robotics, 7*(1), 117–127. https://doi.org/10.1007/s12369-014–0263-x

Albrecht, S. (2023). *ChatGPT und andere Computermodelle zur Sprachverarbeitung – Grundlagen, Anwendungspotenziale und mögliche Auswirkungen* (TAB-Hintergrundpapier Nr. 26). Büro für Technikfolgenabschätzung beim Deutschen Bundestag (TAB). https://doi.org/10.5445/IR/1000158070

American Psychological Association. (2020). *Publication manual of the American Psychological Association (7th ed.).* https://doi.org/10.1037/0000165-000

Anderson, N., Belavy, D. L., Perle, S. M., Hendricks, S., Hespanhol, L., Verhagen, E., & Memon, A. R. (2023). AI did not write this manuscript, or did it? Can we trick the AI text detector into generated texts? The potential future of ChatGPT and AI in Sports & Exercise Medicine manuscript generation. *BMJ Open Sport & Exercise Medicine, 9*(1), Article e001568. https://doi.org/10.1136/bmjsem-2023-001568

Arnold, T.O. (2024). Herausforderungen in der Forschung: Mangelnde Reproduzierbarkeit und Erklärbarkeit. In G. Schreiber & L. Ohly (Eds.), *KI:Text: Diskurse über KI-Textgeneratoren* (pp. 67–80). De Gruyter. https://doi.org/10.1515/9783111351490-006

Arora, R. K., Wei, J., Hicks, R. S., Bowman, P., Quiñonero-Candela, J., Tsimpourlas, F., Sharman, M., Shah, M., Vallone, A., Beutel, A., Heidecke, J., & Singhal, K. (2025). *HealthBench: Evaluating large language models towards improved human health* (No. arXiv:2505.08775). arXiv. https://doi.org/10.48550/arXiv.2505.08775

Banh, L., & Strobel, G. (2023). Generative artificial intelligence. *Electronic Markets, 33*(1), Article 63. https://doi.org/10.1007/s12525-023–00680-1

Baron, R. (2024). AI editing: Are we there yet? *Science Editor, 47*(3), 78–82. https://doi.org/10.36591/SE-4703-18

Barrot, J.S. (2023). Using ChatGPT for second language writing: Pitfalls and potentials. *Assessing Writing, 57,* Article 100745. https://doi.org/10.1016/j.asw.2023.100745

Bašić, Ž., Banovac, A., Kružić, I., & Jerković, I. (2023). ChatGPT-3.5 as writing assistance in students' essays. *Humanities and Social Sciences Communications, 10*(1), Article 750. https://doi.org/10.1057/s41599-023–02269-7

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences, 122*(26), Article e2422633122. https://doi.org/10.1073/pnas.2422633122

Becker, J. (2024). Können Chatbots Romane schreiben? Der Einfluss von KI auf kreatives Schreiben und Erzählen: Eine literaturwissenschaftliche Bestandsaufnahme. In G. Schreiber & L. Ohly (Eds.), *KI:Text: Diskurse über KI-Textgeneratoren* (pp. 83–100). De Gruyter. https://doi.org/10.1515/9783111351490-007

Becker, J., Rush, N., Barnes, B., & Rein, D. (2025). *Measuring the impact of early-2025 AI on experienced open-source developer productivity* (No. arXiv:2507.09089). arXiv. https://doi.org/10.48550/arXiv.2507.09089

Bendel, O. (2024). KI-basierte Textgeneratoren aus Sicht der Ethik. In G. Schreiber & L. Ohly (Eds.), *KI:Text: Diskurse über KI-Textgeneratoren* (pp. 291–306). De Gruyter. https://www.doi.org/10.1515/9783111351490-018

Bender, E.M. (2024). Resisting dehumanization in the age of "AI." *Current Directions in Psychological Science, 33*(2), 114–120. https://doi.org/10.1177/09637214231217286

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency,* 610–623. https://doi.org/10.1145/3442188.3445922

Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrell, T., Harari, Y. N., Zhang, Y.-Q., Xue, L., Shalev-Shwartz, S., Hadfield, G., Clune, J., Maharaj, T., Hutter, F., Baydin, A. G., McIlraith, S., Gao, Q., Acharya, A., Krueger, D., … Mindermann, S. (2024). Managing extreme AI risks amid rapid progress. *Science, 384*(6698), 842–845. https://doi.org/10.1126/science.adn0117

Benichou, L., & ChatGPT. (2023). The role of using ChatGPT AI in writing medical scientific articles. *Journal of Stomatology, Oral and Maxillofacial Surgery, 124*(5), Article 101456. https://doi.org/10.1016/j.jormas.2023.101456

Borji, A. (2023). *A categorical archive of ChatGPT failures* (No. arXiv:2302.03494). arXiv. https://doi.org/10.48550/arXiv.2302.03494

Boussioux, L., Lane, J. N., Zhang, M., Jacimovic, V., & Lakhani, K.R. (2024). The crowdless future? Generative AI and creative problem-solving. *Organization Science, 35*(5), 1589–1607. https://doi.org/10.1287/orsc.2023.18430

Bräuer, G., Hollosi-Boiger, C., Lechleitner, R., & Kreitz, D. (2023). *Literacy Management als Schlüsselkompetenz in einer digitalisierten Welt: Ein Arbeitsbuch für Schreibende, Lehrende und Studierende*. Verlag Barbara Budrich. https://doi.org/10.2307/jj.7762633

Brommer, S., Berendes, J., Bohle-Jurok, U., Buck, I., Girgensohn, K., Grieshammer, E., Gröner, C., Gürtl, F., Hollosi-Boiger, C., Klamm, C., Knorr, D., Limburg, A., Mundorf, M., Stahlberg, N., & Unterpertinger, E. (2023). *Wissenschaftliches Schreiben im Zeitalter von KI gemeinsam verantworten* (Diskussionpaper Nr. 27). Hochschulforum Digitalisierung. https://hochschulforumdigitalisierung.de/wp-content/uploads/2023/11/HFD_DP_27_Schreiben_KI.pdf

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., hÉigeartaigh, S. Ó., Beard, S. J., Belfield, H., Farquhar, S., … Amodei, D. (2018). *The malicious use of artificial intelligence: Forecasting, prevention, and mitigation* (No. arXiv:1802.07228). arXiv. https://doi.org/10.4850/arXiv.1802.07228

Bsharat, S. M., Myrzakhan, A., & Shen, Z. (2024). *Principled instructions are all you need for questioning LLaMA-1/2, GPT-3.5/4* (No. arXiv:2312.16171). arXiv. https://doi.org/10.48550/arXiv.2312.16171

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). *Sparks of artificial general intelligence: Early experiments with GPT-4* (No. arXiv:2303.12712). arXiv. https://doi.org/10.48550/arXiv.2303.12712

Buck, I. (2025). *Wissenschaftliches Schreiben mit KI*. utb. https://doi.org/10.36198/9783838563657

Buck, I., & Limburg, A. (2024). KI und Kognition im Schreibprozess: Prototypen und Implikationen. *JoSch – Journal Für Schreibwissenschaft, 15*(26), 8–23. https://doi.org/10.3278/JOS2401W002

Buck, I., & Wessels, D. (2025). Gut geführt = gut geschrieben? AI Leadership als relevante Kompetenz in der Kollaboration mit KI-Tools. In G. Brägger & H.-G. Rolff (Eds.), *Handbuch Lernen mit digitalen Medien. Wege der Transformation* (3. Auflage, pp. 863–880). Beltz Verlag.

Cahyawijaya, S., Chen, D., Bang, Y., Khalatbari, L., Wilie, B., Ji, Z., Ishii, E., & Fung, P. (2025). High-dimension human value representation in large language models. In L. Chiruzzo, A. Ritter, & L. Wang (Eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 5303–5330). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-long.274

Casal, J. E., & Kessler, M. (2023). Can linguists distinguish between ChatGPT/AI and human writing?: A study of research ethics and academic publishing. *Research Methods in Applied Linguistics, 2*(3), Article 100068. https://doi.org/10.1016/j.rmal.2023.100068

Center for AI Safety. (2023). *Statement on AI risk*. https://safe.ai/work/statement-on-ai-risk

Chang, C.-Y., Lin, H.-C., Yin, C., & Yang, K.-H. (2025). Generative AI-assisted reflective writing for improving students' higher order thinking: Evidence from quantitative and epistemic network analysis. *Educational Technology & Society, 28*(1), 270–285. https://doi.org/10.30191/ETS.202501_28(1).TP03

Chen, C., & Shu, K. (2024). Combating misinformation in the age of LLMs: Opportunities and challenges. *AI Magazine, 45*(3), 354–368. https://doi.org/10.1002/aaai.12188

Cheung, M. (2024). *A reality check of the benefits of LLM in business* (No. arXiv:2406.10249). arXiv. https://doi.org/10.48550/arXiv.2406.10249

Chiang, T. (2023, February 9). ChatGPT Is a blurry JPEG of the web. *The New Yorker*. https://www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web

Chiriatti, M., Ganapini, M., Panai, E., Ubiali, M., & Riva, G. (2024). The case for human–AI interaction as system 0 thinking. *Nature Human Behaviour, 8*(10), 1829–1830. https://doi.org/10.1038/s41562-024–01995-5

Crawford, K. (2024). *Atlas der KI: Die materielle Wahrheit hinter den neuen Datenimperien* (F. Lachmann, Trans.). C.H. Beck. https://doi.org/10.17104/9783406823343

Curtis, N., & ChatGPT. (2023). To ChatGPT or not to ChatGPT? The impact of artificial intelligence on academic publishing. *The Pediatric Infectious Disease Journal, 42*(4), 275. https://doi.org/10.1097/INF.0000000000003852

Dathathri, S., See, A., Ghaisas, S., Huang, P.-S., McAdam, R., Welbl, J., Bachani, V., Kaskasoli, A., Stanforth, R., Matejovicova, T., Hayes, J., Vyas, N., Merey, M. A., Brown-Cohen, J., Bunel, R., Balle, B., Cemgil, T., Ahmed, Z., Stacpoole, K., … Kohli, P. (2024). Scalable watermarking for identifying large language model outputs. *Nature, 634*(8035), 818–823. https://doi.org/10.1038/s41586-024–08025-4

Day, T. (2023). A preliminary investigation of fake peer-reviewed citations and references generated by ChatGPT. *The Professional Geographer, 75*(6), 1024–1027. https://doi.org/10.1080/00330124.2023.2190373

Dell'Acqua, F., McFowland III, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K.R. (2023). *Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality* (SSRN Scholarly Paper No. 4573321). Social Science Research Network. https://doi.org/10.2139/ssrn.4573321

Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J.M. (2019). Hybrid intelligence. *Business & Information Systems Engineering, 61*(5), 637–643. https://doi.org/10.1007/s12599-019–00595-2

Deng, R., Jiang, M., Yu, X., Lu, Y., & Liu, S. (2025). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. *Computers & Education, 227,* Article 105224. https://doi.org/10.1016/j.compedu.2024.105224

Deutscher Ethikrat. (2023, March 20). *Mensch und Maschine – Herausforderungen durch Künstliche Intelligenz (Stellungnahme).* https://www.ethikrat.org/fileadmin/Publikationen/Stellungnahmen/deutsch/stellungnahme-mensch-und-maschine.pdf

Dobrin, S.I. (2023). *AI and writing.* Broadview Press.

Dornis, T. W., & Stober, S. (2024). *Urheberrecht und Training generativer KI-Modelle: Technologische und juristische Grundlagen.* Nomos. https://doi.org/10.5771/9783748949558

Doshi, A. R., & Hauser, O.P. (2024). Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances, 10*(28), Article eadn5290. https://doi.org/10.1126/sciadv.adn5290

Dowling, M., & Lucey, B. (2023). ChatGPT for (finance) research: The Bananarama conjecture. *Finance Research Letters, 53,* Article 103662. https://doi.org/10.1016/j.frl.2023.103662

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., … Wright, R. (2023). Opinion paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management, 71,* Article 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Ebner, P., & Szczuka, J. (2025). *Predicting romantic human-chatbot relationships: A mixed-method study on the key psychological factors* (No. arXiv:2503.00195). arXiv. https://doi.org/10.48550/arXiv.2503.00195

Eichenberger, A., Thielke, S., & Van Buskirk, A. (2025). A case of bromism Influenced by use of artificial intelligence. *Annals of Internal Medicine: Clinical Cases, 4*(8), Article e241260. https://doi.org/10.7326/aimcc.2024.1260

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review, 114*(4), 864–886. https://doi.org/10.1037/0033-295X.114.4.864

Errica, F., Siracusano, G., Sanvito, D., & Bifulco, R. (2025). What did I do wrong? Quantifying LLMs' sensitivity and consistency to prompt engineering. In L. Chiruzzo, A. Ritter, & L. Wang (Eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 1543–1558). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-long.73

European Union Agency for Fundamental Rights. (2022). *Bias in algorithms: Artificial intelligence and discrimination.* https://fra.europa.eu/en/publication/2022/bias-algorithm

Everett, S. S., Bunning, B. J., Jain, P., Lopez, I., Agarwal, A., Desai, M., Gallo, R., Goh, E., Kadiyala, V. B., Kanjee, Z., Koshy, J. M., Olson, A., Rodman, A., Schulman, K., Strong, E., Chen, J. H., & Horvitz, E. (2025). *From tool to teammate: A randomized controlled trial of clinician-AI collaborative workflows for diagnosis* (No. medRxiv:2025.06.07.25329176). medRxiv. https://doi.org/10.1101/2025.06.07.25329176

Fang, A., & Perkins, J. (2024). Large language models (LLMs): Risks and policy implications. *MIT Science Policy Review, 5,* 134–145. https://doi.org/10.38105/spr.3qrco9kp8x

Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2024). A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education and Teaching International, 61*(3), 460–474. https://doi.org/10.1080/14703297.2023.2195846

Ferrara, E. (2023a). *Should ChatGPT be biased? Challenges and risks of bias in large language models* (No. arXiv:2304.03738). arXiv. https://doi.org/10.48550/arXiv.2304.03738

Ferrara, E. (2023b, July 19). *Eliminating bias in AI may be impossible – a computer scientist explains how to tame it instead.* The Conversation. http://theconversation.com/eliminating-bias-in-ai-may-be-impossible-a-computer-scientist-explains-how-to-tame-it-instead-208611

Ferretti, T. (2022). An institutionalist approach to AI ethics: Justifying the priority of government regulation over self-regulation. *Moral Philosophy and Politics, 9*(2), 239–265. https://doi.org/10.1515/mopp-2020-0056

Filipović, A., Burchardt, A., Michel, A., Puzio, A., Reinmann, G., Schaumann, P., Tippe, U., Wan, M., & Wilder, N. (2025). *Künstliche Intelligenz: Grundlagen für das Handeln in der Hochschullehre* (Arbeitspapier Nr. 86). Hochschulforum Digitalisierung. https://hochschulforumdigitalisierung.de/wp-content/uploads/2025/01/HFD_AP_86_Kuenstliche-Intelligenz_Grundlagen-fuer-das-Handeln.pdf

Fjelland, R. (2020). Why general artificial intelligence will not be realized. *Humanities and Social Sciences Communications, 7*(1), 1–9. https://doi.org/10.1057/s41599-020–0494-4

Fleckenstein, J., Meyer, J., Jansen, T., Keller, S. D., Köller, O., & Möller, J. (2024). Do teachers spot AI? Evaluating the detectability of AI-generated texts among student essays. *Computers and Education: Artificial Intelligence, 6,* Article 100209. https://doi.org/10.1016/j.caeai.2024.100209

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences, 3*(4), 128–135. https://doi.org/10.1016/S 1364–6613(99)01294-2

Friedrich, J.-D., Tobor, J., & Wan, M. (2024). *9 Mythen über generative KI in der Hochschulbildung* (Diskussionpaper Nr. 29). Hochschulforum Digitalisierung. https://hochschulforumdigitalisierung.de/wp-content/uploads/2024/02/9_Mythen_ueber_generative_KI_in_der_Hochschulbildung.pdf

Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2023). Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *Npj Digital Medicine, 6,* Article 75. https://doi.org/10.1038/s41746-023–00819-6

Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies, 15*(1), Article 6. https://doi.org/10.3390/soc15010006

Gibbs, J. (2025, July 27). *The only ChatGPT prompt that matters.* Medium. https://medium.com/@jordan_gibbs/the-only-chatgpt-prompt-that-matters-f16dd85e43b3

Gimpel, H., Hall, K., Decker, S., Eymann, T., Lämmermann, L., Mädche, A., Röglinger, R., Ruiner, C., Schoch, M., Schoop, M., Urbach, N., & Vandirk, S. (2023). *Unlocking the power of generative AI models and systems such as GPT-4 and ChatGPT for higher education: A guide for students and lecturers.* University of Hohenheim.

Giray, L. (2023). Prompt engineering with ChatGPT: A guide for academic writers. *Annals of Biomedical Engineering, 51*(12), 2629–2633. https://doi.org/10.1007/s10439-023–03272-4

Goedecke, S. (2025, May 6). *I don't care about your magic prompts.* Seangoedecke.Com. https://www.seangoedecke.com/magic-prompts/

Goertzel, B. (2014). Artificial general intelligence: Concept, state of the art, and future prospects. *Journal of Artificial General Intelligence, 5*(1), 1–48. https://doi.org/10.2478/jagi-2014-0001

Gredel, E., Pospiech, U., & Schindler, K. (2024). Künstliche Intelligenz und Schreiben in (hoch-) schulischen Kontexten. *Zeitschrift für germanistische Linguistik, 52*(2), 378–404. https://doi.org/10.1515/zgl-2024-2018

Heersmink, R. (2024). Use of large language models might affect our cognitive skills. *Nature Human Behaviour, 8*(5), 805–806. https://doi.org/10.1038/s41562-024–01859-y

Heikkilä, M., & Honan, M. (2024, October 31). *OpenAI brings a new web search tool to ChatGPT.* MIT Technology Review. https://www.technologyreview.com/2024/10/31/1106472/chatgpt-now-lets-you-search-the-internet/

Hicks, M. T., Humphries, J., & Slater, J. (2024). ChatGPT is bullshit. *Ethics and Information Technology, 26*(2), Article 38. https://doi.org/10.1007/s10676-024–09775-5

Hirsch, N. (2023, October 4). KI-Strategie: Prompting, Referenzieren & Regeln. *eBildungslabor.* https://ebildungslabor.de/blog/ki-strategie-prompting-referenzieren-regeln/

Höfler, E. (2024). Digitale Rolle rückwärts in Schweden – eine Bilanz. *On. Lernen in Der Digitalen Welt, 19,* 28–29.

Huang, J., & Tan, M. (2023). The role of ChatGPT in scientific communication: Writing better scientific review articles. *American Journal of Cancer Research, 13*(4), 1148–1154.

Hutson, J. (2025). Human-AI collaboration in writing: A multidimensional framework for creative and intellectual authorship. *International Journal of Changes in Education.* Advance online publication. https://doi.org/10.47852/bonviewIJCE52024908

Ingley, S. J., & Pack, A. (2023). Leveraging AI tools to develop the writer rather than the writing. *Trends in Ecology & Evolution, 38*(9), 785–787. https://doi.org/10.1016/j.tree.2023.05.007

Jahnel, J., & Heil, R. (2024). KI-Textgeneratoren als soziotechnisches Phänomen: Ansätze zur Folgenabschätzung und Regulierung. In G. Schreiber & L. Ohly (Eds.), *KI: Text: Diskurse über KI-Textgeneratoren* (pp. 341–354). De Gruyter. https://doi.org/10.1515/9783111351490-021

Jiang, F., & Hyland, K. (2025). Does ChatGPT argue like students? Bundles in argumentative essays. *Applied Linguistics, 46*(3), 375–391. https://doi.org/10.1093/applin/amae052

Jordan, M. I., & Mitchell, T.M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Kacena, M. A., Plotkin, L. I., & Fehrenbacher, J. C. (2024). The use of artificial intelligence in writing scientific review articles. *Current Osteoporosis Reports, 22*(1), 115–121. https://doi.org/10.1007/s11914-023–00852-0

Karamolegkou, A., Li, J., Zhou, L., & Søgaard, A. (2023). Copyright violations and large language models. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 7403–7412). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.458

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103,* Article 102274. https://doi.org/10.1016/j.lindif.2023.102274

Katz, D. M., Bommarito, M. J., Gao, S., & Arredondo, P. (2023). *GPT-4 passes the bar exam* (SSRN Scholarly Paper No. 4389233). Social Science Research Network. https://doi.org/10.2139/ssrn.4389233

Khalifa, M., & Albadawy, M. (2024). Using artificial intelligence in academic writing and research: An essential productivity tool. *Computer Methods and Programs in Biomedicine Update, 5,* Article 100145. https://doi.org/10.1016/j.cmpbup.2024.100145

Kim, S.-J. (2023). Trends in research on ChatGPT and adoption-related issues discussed in articles: A narrative review. *Science Editing, 11*(1), 3–11. https://doi.org/10.6087/kcse.321

Klein, O. (2025, May 13). *ChatGPT und Co: KI-Modelle erzählen immer mehr Unsinn.* ZDFheute. https://www.zdfheute.de/panorama/kuenstliche-intelligenz-ki-chatgpt-sprachmodelle-halluzinationen-entwicklung-100.html

Kobak, D., González-Márquez, R., Horvát, E.-Á., & Lause, J. (2025). *Delving into Chat GPT usage in academic writing through excess vocabulary* (No. arXiv:2406.07016). arXiv. https://doi.org/10.48550/arXiv.2406.07016

Koch, G. (2025, May 22). *Reclaiming your words: Fighting stealth watermarks in AI-generated text & why it matters (a developer's perspective).* Hugging Face. https://huggingface.co/blog/cronos3k/fighting-stealth-watermarks-in-ai-generated-text-w

Konrad, A., & Cai, K. (2023, February 2). *Inside ChatGPT's breakout moment and the race to put AI to work*. Forbes. https://www.forbes.com/sites/alexkonrad/2023/02/02/inside-chatggpts-breakout-moment-and-the-race-for-the-future-of-ai/

Kosberg, R. (2024). *Write with AI: Conquer writer's block, unleash your creativity, and write your book using artificial intelligence*. Best Seller Publishing.

Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X.-H., Beresnitzky, A. V., Braunstein, I., & Maes, P. (2025). *Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task* (No. arXiv:2506.08872). arXiv. https://doi.org/10.48550/arXiv.2506.08872

Krüger, N. (2023). ChatGPT et al. Was bedeutet ChatGPT für den wissenschaftlichen Schreibprozess? *Exposé – Zeitschrift Für Wissenschaftliches Schreiben Und Publizieren, 4*(2), 12–15. https://doi.org/10.3224/expose.v4i2.03

Krupp, L., Steinert, S., Kiefer-Emmanouilidis, M., Avila, K. E., Lukowicz, P., Kuhn, J., Küchemann, S., & Karolus, J. (2024). Unreflected acceptance – Investigating the negative consequences of ChatGPT-assisted problem solving in physics education. In F. Lorig, J. Tucker, A. Dahlgren Lindström, F. Dignum, P. Murukannaiah, A. Theodorou, & P. Yolum (Eds.), *HHAI 2024: Hybrid Human AI Systems for the Social Good. Proceedings of the Third International Conference on Hybrid Human-Artificial Intelligence* (pp. 199–212). IOS Press. https://doi.org/10.3233/FAIA240195

Kshetri, N. (2023). Cybercrime and privacy threats of large language models. *IT Professional, 25*(3), 9–13. https://doi.org/10.1109/MITP.2023.3275489

Kultusministerkonferenz der Bundesrepublik Deutschland. (2025). *Handlungsempfehlung für die Bildungsverwaltung zum Umgang mit Künstlicher Intelligenz in schulischen Bildungsprozessen.* https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2024/2024_10_10-Handlungsempfehlung-KI.pdf

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., Leon, L. D., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health, 2*(2), Article e0000198. https://doi.org/10.1371/journal.pdig.0000198

Laban, P., Hayashi, H., Zhou, Y., & Neville, J. (2025). *LLMs get lost in multi-turn conversation* (No. arXiv:2505.06120). arXiv. https://doi.org/10.48550/arXiv.2505.06120

Laupichler, M. C., Aster, A., Schirch, J., & Raupach, T. (2022). Artificial intelligence literacy in higher and adult education: A scoping literature review. *Computers and Education: Artificial Intelligence, 3,* Article 100101. https://doi.org/10.1016/j.caeai.2022.100101

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lee, B. C., & Chung, J. (2024). An empirical investigation of the impact of ChatGPT on creativity. *Nature Human Behaviour, 8*(10), 1906–1914. https://doi.org/10.1038/s41562-024–01953-1

Lee, H.-P., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., & Wilson, N. (2025). The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In N. Yamashita, V. Evers, K. Yatani, X. Ding, B. Lee, M. Chetty, & P. Toups-Dugas (Eds.), *CHI '25: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (p. Article 1121). Association for Computing Machinery. https://doi.org/10.1145/3706598.3713778

Lehmann, M., Cornelius, P. B., & Sting, F. J. (2025). *AI meets the classroom: When do large language models harm learning?* (SSRN Scholarly Paper No. 4941259). Social Science Research Network. https://doi.org/10.2139/ssrn.4941259

Li, C., Wang, J., Zhang, Y., Zhu, K., Hou, W., Lian, J., Luo, F., Yang, Q., & Xie, X. (2023). *Large language models understand and can be enhanced by emotional stimuli* (No. arXiv:2307.11760). arXiv. https://doi.org/10.48550/arXiv.2307.11760

Limburg, A., Bohle-Jurok, U., Buck, I., Grieshammer, E., Gröpler, J., Knorr, D., Mundorf, M., Schindler, K., & Wilder, N. (2023). *Zehn Thesen zur Zukunft des wissenschaftlichen Schreibens* (Diskussionpaper Nr. 23). Hochschulforum Digitalisierung. https://hochschulforumdigitalisierung.de/wp-content/uploads/2023/09/HFD_DP_23_Zukunft_Schreiben_Wissenschaft.pdf

Lin, S., Hilton, J., & Evans, O. (2022). *TruthfulQA: Measuring how models mimic human falsehoods* (No. arXiv:2109.07958). arXiv. https://doi.org/10.48550/arXiv.2109.07958

Lin, Z. (2024a). How to write effective prompts for large language models. *Nature Human Behaviour, 8*(4), 611–615. https://doi.org/10.1038/s41562-024–01847-2

Lin, Z. (2024b). Techniques for supercharging academic writing with generative AI. *Nature Biomedical Engineering, 9,* 426–431. https://doi.org/10.1038/s41551-024–01185-8

Lingard, L. (2023). Writing with ChatGPT: An illustration of its capacity, limitations & implications for academic writers. *Perspectives on Medical Education, 12*(1), 261–270. https://doi.org/10.5334/pme.1072

Liu, Y., Cao, J., Liu, C., Ding, K., & Jin, L. (2024). *Datasets for large language models: A comprehensive survey* (No. arXiv:2402.18041). arXiv. https://doi.org/10.48550/arXiv.2402.18041

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems,* Paper 598. https://doi.org/10.1145/3313831.3376727

Lu, D. (2025, January 28). We tried out DeepSeek. It worked well, until we asked it about Tiananmen Square and Taiwan. *The Guardian.* https://www.theguardian.com/technology/2025/jan/28/we-tried-out-deepseek-it-works-well-until-we-asked-it-about-tiananmen-square-and-taiwan

Luccioni, A. S., Pistilli, G., Sefala, R., & Moorosi, N. (2025). *Bridging the gap: Integrating ethics and environmental sustainability in AI research and practice* (No. arXiv:2504.00797). arXiv. https://doi.org/10.48550/arXiv.2504.00797

Luo, Y., Yang, Z., Meng, F., Li, Y., Zhou, J., & Zhang, Y. (2025). *An empirical study of catastrophic forgetting in large language models during continual fine-tuning* (No. arXiv:2308.08747). arXiv. https://doi.org/10.48550/arXiv.2308.08747

Maggiori, E. (2023). *Smart until it's dumb: Why artificial intelligence keeps making epic mistakes (and why the AI bubble will burst).* Applied Maths Ltd.

Marji, Z., & Licato, J. (2025). Evaluating large language models' ability to generate interpretive arguments. *Argument & Computation, 16*(3), 362–404. https://doi.org/10.3233/aac-230014

Markey, B., Brown, D. W., Laudenbach, M., & Kohler, A. (2024). Dense and disconnected: Analyzing the sedimented style of ChatGPT-generated text at scale. *Written Communication, 41*(4), 571–600. https://doi.org/10.1177/07410883241263528

Meier-Vieracker, S. (2024). Uncreative Academic Writing: Sprachtheoretische Überlegungen zu Künstlicher Intelligenz in der akademischen Textproduktion. In G. Schreiber & L. Ohly (Eds.), *KI:Text: Diskurse über KI-Textgeneratoren* (pp. 133–144). De Gruyter. https://doi.org/10.1515/9783111351490-010

Meincke, L., Mollick, E. R., Mollick, L., & Shapiro, D. (2025). *Prompting science report 1: Prompt engineering is complicated and contingent* (SSRN Scholarly Paper No. 5165270). Social Science Research Network. https://doi.org/10.2139/ssrn.5165270

Meincke, L., Mollick, E. R., & Terwiesch, C. (2024). *Prompting diverse ideas: Increasing AI idea variance* (SSRN Scholarly Paper No. 4708466). Social Science Research Network. https://doi.org/10.2139/ssrn.4708466

Merken, S. (2024, November 27). Texas lawyer fined for AI use in latest sanction over fake citations. *Reuters.* https://www.reuters.com/legal/government/texas-lawyer-fined-ai-use-latest-sanction-over-fake-citations-2024-11-26/

Messeri, L., & Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature, 627*(8002), 49–58. https://doi.org/10.1038/s41586-024–07146-0

Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A., & Fleckenstein, J. (2024). Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence, 6,* Article 100199. https://doi.org/10.1016/j.caeai.2023.100199

Mikros, G. (2025). Beyond the surface: Stylometric analysis of GPT-4o's capacity for literary style imitation. *Digital Scholarship in the Humanities, 40*(2), 587–600. https://doi.org/10.1093/llc/fqaf035

Milmo, D. (2024, February 22). Google pauses AI-generated images of people after ethnicity criticism. *The Guardian.* https://www.theguardian.com/technology/2024/feb/22/google-pauses-ai-generated-images-of-people-after-ethnicity-criticism

Mollick, E. (2023a, March 29). *How to use AI to do practical stuff: A new guide.* One Useful Thing. https://www.oneusefulthing.org/p/how-to-use-ai-to-do-practical-stuff

Mollick, E. (2023b, April 26). *A guide to prompting AI (for what it is worth).* One Useful Thing. https://www.oneusefulthing.org/p/a-guide-to-prompting-ai-for-what

Mollick, E. (2023c, November 1). *Working with AI: Two paths to prompting.* One Useful Thing. https://www.oneusefulthing.org/p/working-with-ai-two-paths-to-prompting

Mollick, E. (2024a, August 30). *Post-apocalyptic education.* One Useful Thing. https://www.oneusefulthing.org/p/post-apocalyptic-education

Mollick, E. (2024b, November 24). *Getting started with AI: Good enough prompting.* One Useful Thing. https://www.oneusefulthing.org/p/getting-started-with-ai-good-enough

Mollick, E. (2025, July 7). *Against "brain damage."* One Useful Thing. https://www.oneusefulthing.org/p/against-brain-damage

Monckton, V. (2025). Has AI replaced editors? *Valerie Monckton.* https://www.valeriemonckton.com/has-ai-replaced-editors/

Montes, C. M., & Khojah, R. (2025). *Emotional strain and frustration in LLM interactions in software engineering* (No. arXiv:2504.10050). arXiv. https://doi.org/10.48550/arXiv.2504.10050

Montgomerie, A. (2024, November). *Editor vs AI: Can grammarly edit itself?* Right Angels and Polo Bears: Adventures in Editing. https://scieditor.ca/2024/11/editor-vs-ai-can-grammarly-edit-itself/

Moravec, H. (1988). *Mind children: The future of robot and human intelligence.* Harvard University Press.

Morrison, A. (2023). Meta-writing: AI and writing. *Composition Studies, 51*(1), 155–161. https://compstudiesjournal.com/wp-content/uploads/2023/06/morrison.pdf

Naddaf, M. (2025). AI linked to explosion of low-quality biomedical research papers. *Nature, 641*(8065), 1080–1081. https://doi.org/10.1038/d41586-025–01592-0

Navigli, R., Conia, S., & Ross, B. (2023). Biases in large language models: Origins, inventory, and discussion. *Journal of Data and Information Quality, 15*(2), Article 10. https://doi.org/10.1145/3597307

Neumeister, L. (2023, June 9). *Lawyers blame ChatGPT for tricking them into citing bogus case law.* Associated Press News. https://apnews.com/article/artificial-intelligence-chatgpt-courts-e15023d7e6fdf4f099aa122437dbb59b

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence, 2,* Article 100041. https://doi.org/10.1016/j.caeai.2021.100041

Nguyen, A., Hong, Y., Dang, B., & Huang, X. (2024). Human-AI collaboration patterns in AI-assisted academic writing. *Studies in Higher Education, 49*(5), 847–864. https://doi.org/10.1080/03075079.2024.2323593

Niloy, A. C., Akter, S., Sultana, N., Sultana, J., & Rahman, S. I. U. (2024). Is Chatgpt a menace for creative writing ability? An experiment. *Journal of Computer Assisted Learning, 40*(2), 919–930. https://doi.org/10.1111/jcal.12929

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science, 381*(6654), 187–192. https://doi.org/10.1126/science.adh2586

Oertner, M. (2024). ChatGPT als Recherchetool? Fehlertypologie, technische Ursachenanalyse und hochschuldidaktische Implikationen. *Bibliotheksdienst, 58*(5), 259–297. https://doi.org/10.1515/bd-2024-0042

Ooi, K.-B., Tan, G. W.-H., Al-Emran, M., Al-Sharafi, M. A., Capatina, A., Chakraborty, A., Dwivedi, Y. K., Huang, T.-L., Kar, A. K., Lee, V.-H., Loh, X.-M., Micu, A., Mikalef, P., Mogaji, E., Pandey, N., Raman, R., Rana, N. P., Sarker, P., Sharma, A., … Wong, L.-W. (2025). The potential of generative artificial intelligence across disciplines: Perspectives and future directions. *Journal of Computer Information Systems, 65*(1), 76–107. https://doi.org/10.1080/08874417.2023.2261010

OpenAI. (2025a). *OpenAI o3 and o4-mini system card.* https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf

OpenAI. (2025b, April 29). *Europe terms of use.* OpenAI. https://openai.com/policies/eu-terms-of-use

OpenAI. (2025c, May 2). *Expanding on what we missed with sycophancy.* OpenAI. https://openai.com/index/expanding-on-sycophancy

Patel, N., Nagpal, P., Shah, T., Sharma, A., Malvi, S., & Lomas, D. (2023). Improving mathematics assessment readability: Do large language models help? *Journal of Computer Assisted Learning, 39*(3), 804–822. https://doi.org/10.1111/jcal.12776

Paullada, A., Raji, I. D., Bender, E. M., Denton, E., & Hanna, A. (2021). Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns, 2*(11), Article 100336. https://doi.org/10.1016/j.patter.2021.100336

Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., Jones, A., Chen, A., Mann, B., Israel, B., Seethor, B., McKinnon, C., Olah, C., Yan, D., Amodei, D., … Kaplan, J. (2022). *Discovering language model behaviors with model-written evaluations* (No. arXiv:2212.09251). arXiv. https://doi.org/10.48550/arXiv.2212.09251

Peter, S., Riemer, K., & West, J. D. (2025). The benefits and dangers of anthropomorphic conversational agents. *Proceedings of the National Academy of Sciences, 122*(22), Article e2415898122. https://doi.org/10.1073/pnas.2415898122

Peterson, A. J. (2025). AI and the problem of knowledge collapse. *AI & SOCIETY, 40,* 3249–3269. https://doi.org/10.1007/s00146-024–02173-x

Placani, A. (2024). Anthropomorphism in AI: Hype and fallacy. *AI and Ethics, 4*(3), 691–698. https://doi.org/10.1007/s43681-024–00419-4

Poorghadiri, S. (2025, February 13). *Alternative Hilfe per App: Junge Menschen entdecken KI als Therapie-Ersatz – Schweizer Psychologen skeptisch.* Tages-Anzeiger. https://www.tagesanzeiger.ch/psychotherapie-ist-kuenstliche-intelligenz-der-neue-therapiersatz-999616034861

Rafner, J., Dellermann, D., Hjorth, A., Verasztó, D., Kampf, C., Mackay, W., & Sherson, J. (2021). Deskilling, upskilling, and reskilling: A case for hybrid intelligence. *Morals & Machines, 1*(2), 24–39. https://doi.org/doi.org/10.5771/2747–5174-2021-2-24

Rankin, J. (2025, February 19). EU accused of leaving 'devastating' copyright loophole in AI Act. *The Guardian.* https://www.theguardian.com/technology/2025/feb/19/eu-accused-of-leaving-devastating-copyright-loophole-in-ai-act

Ray, S. (2023, May 2). *Samsung bans ChatGPT among employees after sensitive code leak.* Forbes. https://www.forbes.com/sites/siladityaray/2023/05/02/samsung-bans-chatgpt-and-other-chatbots-for-employees-after-sensitive-code-leak/

Rebala, G., Ravi, A., & Churiwala, S. (2019). *An introduction to machine learning.* Springer. https://doi.org/10.1007/978-3-030–15729-6

Reinmann, G. (2023). *Deskilling durch Künstliche Intelligenz?* (Diskussionpaper Nr. 25). Hochschulforum Digitalisierung. https://hochschulforumdigitalisierung.de/sites/default/files/dateien/HFD_DP_25_Deskilling.pdf

Rezaee, O. (2025, August 1). ChatGPT-Leak: Wie Tausende ChatGPT-Gespräche öffentlich wurden. *Die Zeit.* https://www.zeit.de/digital/internet/2025-08/chatgpt-leak-google-chats-oeffentlich-privat

Russell, S., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (global edition, 4th ed.). Pearson.

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2025). *Can AI-generated text be reliably detected?* (No. arXiv:2303.11156). arXiv. https://doi.org/10.48550/arXiv.2303.11156

Saha, A. (1998). Technological innovation and Western values. *Technology in Society, 20*(4), 499–520. https://doi.org/10.1016/S 0160-791X(98)00030-X

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2025). *A systematic survey of prompt engineering in large language models: Techniques and applications* (No. arXiv:2402.07927). arXiv. https://doi.org/10.48550/arXiv.2402.07927

Salden, P., & Leschke, J. (2023). *Didaktische und rechtliche Perspektiven auf KI-gestütztes Schreiben in der Hochschulbildung.* Ruhr-Universität Bochum. https://doi.org/10.13154/294–9734

Salvagno, M., Taccone, F. S., & Gerli, A.G. (2023). Can artificial intelligence help for scientific writing? *Critical Care, 27,* Article 75. https://doi.org/10.1186/s13054-023–04380-2

Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Müller, K.-R. (2021). Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE, 109*(3), 247–278. https://doi.org/10.1109/JPROC.2021.3060483

Sarker, I.H. (2021). Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions. *SN Computer Science, 2*(6), Article 420. https://doi.org/10.1007/s42 979-021–00815-1

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., Tow, J., Rush, A. M., Biderman, S., Webson, A., Ammanamanchi, P. S., Wang, T., Sagot, B., Muennighoff, N., Villanova del Moral, A., … Wolf, T. (2023). *BLOOM: A 176B-parameter open-access multilingual language model* (No. arXiv:2211.05100). arXiv. https://doi.org/10.4855 0/arXiv.2211.05100

Schicker, S., & Akbulut, M. (2023). ChatGPT – maschinelle und menschliche Textsortenkompetenz. In S. Schicker & L. Miškulin Saletović (Eds.), *Sprachliche Handlungsmuster & Text(sorten)kompetenz* (pp. 169–196). https://doi.org/10.25364/978390337426311

Schlatter, G. (2005). Animismus. In C. Auffarth, J. Bernard, H. Mohr, A. Imhof, & S. Kurre (Eds.), *Metzler Lexikon Religion: Gegenwart—Alltag—Medien* (p. 61). J.B. Metzler. https://doi.org/10.10 07/978-3-476-00091-0_19

Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., Li, Y., Gupta, A., Han, H., Schulhoff, S., Dulepet, P. S., Vidyadhara, S., Ki, D., Agrawal, S., Pham, C., Kroiz, G., Li, F., Tao, H., Srivastava, A., … Resnik, P. (2025). *The prompt report: A systematic survey of prompt engineering techniques* (No. arXiv:2406.06608). arXiv. https://doi.org/10.48550/arXiv.2406.06608

Schwarcz, D., Manning, S., Barry, P., Cleveland, D. R., Prescott, J. J., & Rich, B. (2025). *AI-powered lawyering: AI reasoning models, retrieval augmented generation, and the future of legal practice* (SSRN Scholarly Paper No. 5162111). Social Science Research Network. https://doi.org/10.2139/ ssrn.5162111

Searle, J.R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*(3), 417–424. https://doi.org/10.1017/S0140525X00005756

Shojaee, P., Mirzadeh, I., Alizadeh, K., Horton, M., Bengio, S., & Farajtabar, M. (2025). *The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity.* Apple Machine Learning Research. https://machinelearning.apple.com/research/illu sion-of-thinking

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature, 631*(8022), 755–759. https://doi.or g/10.1038/s41586-024–07566-y

Southworth, J., Migliaccio, K., Glover, J., Glover, J., Reed, D., McCarty, C., Brendemuhl, J., & Thomas, A. (2023). Developing a model for AI across the curriculum: Transforming the higher education landscape via innovation in AI literacy. *Computers and Education: Artificial Intelligence, 4,* Article 100127. https://doi.org/10.1016/j.caeai.2023.100127

Stadler, M., Bannert, M., & Sailer, M. (2024). Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry. *Computers in Human Behavior, 160,* Article 108386. https://doi.org/10.1016/j.chb.2024.108386

Steinert, C. V., & Kazenwadel, D. (2025). How user language affects conflict fatality estimates in ChatGPT. *Journal of Peace Research, 62*(4), 1128–1143. https://doi.org/10.1177/00223433241279381

Sun, Y., Sheng, D., Zhou, Z., & Wu, Y. (2024). AI hallucination: Towards a comprehensive classification of distorted information in artificial intelligence-generated content. *Humanities and Social Sciences Communications, 11*(1), 1–14. https://doi.org/10.1057/s41599-024–03811-x

Sun, Z., Wang, Q., Wang, H., Zhang, X., & Xu, J. (2025). *Detection and mitigation of hallucination in large reasoning models: A mechanistic perspective* (No. arXiv:2505.12886). arXiv. https://doi.org/10.48550/arXiv.2505.12886

Suwała, S., Szulc, P., Guzowski, C., Kamińska, B., Dorobiała, J., Wojciechowska, K., Berska, M., Kubicka, O., Kosturkiewicz, O., Kosztulska, B., Rajewska, A., & Junik, R. (2024). ChatGPT-3.5 passes Poland's medical final examination—Is it possible for ChatGPT to become a doctor in Poland? *SAGE Open Medicine, 12,* 1–7. https://doi.org/10.1177/20503121241257777

Teng, M. F. (2024). "ChatGPT is the companion, not enemies": EFL learners' perceptions and experiences in using ChatGPT for feedback in writing. *Computers and Education: Artificial Intelligence, 7,* Article 100270. https://doi.org/10.1016/j.caeai.2024.100270

Thorp, H. H. (2023). ChatGPT is fun, but not an author. *Science, 379*(6630), 313. https://doi.org/10.1126/science.adg7879

Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. (2023). *Nature, 613*(7945), 612. https://doi.org/10.1038/d41586-023–00191-1

Troendle, S. (2023, December 14). *Kommentar: Axel Springer-Verlag und OpenAI gehen Partnerschaft ein.* SWR Wissen. https://www.swr.de/wissen/kommentar-springer-verlag-kooperation-mit-chat-gpt-kuenstliche-intelligenz-100.html

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, LIX*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

United States Copyright Office. (2025a). *Copyright and artificial intelligence, part 2: Copyrightability.* https://copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-2-Copyrightability-Report.pdf

United States Copyright Office. (2025b). *Copyright and artificial intelligence, part 3: Generative AI training* [Pre-publication version]. https://copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf

Varanasi, L. (2025, April 10). *ChatGPT can now remember everything you ever told it.* Business Insider. https://www.businessinsider.com/chatgpt-memory-remember-everything-you-ever-told-it-2025-4

Varela, I. D., Romero-Sorozabal, P., Rocon, E., & Cebrian, M. (2025). *Rethinking the illusion of thinking* (No. arXiv:2507.01231). arXiv. https://doi.org/10.48550/arXiv.2507.01231

Waldo, J., & Boussard, S. (2024). GPTs and hallucination: Why do large language models hallucinate? *Queue, 22*(4), 19–33. https://doi.org/10.1145/3688007

Wallwork, A. (2024). *AI-assisted writing and presenting in english.* Springer Nature. https://doi.org/10.1007/978-3-031–48147-5

Walter, Y. (2024a). Embracing the future of artificial intelligence in the classroom: The relevance of AI literacy, prompt engineering, and critical thinking in modern education. *International Journal of Educational Technology in Higher Education, 21*(1), Article 15. https://doi.org/10.1186/s41239-024–00448-3

Walter, Y. (2024b). Managing the race to the moon: Global policy and governance in artificial intelligence regulation—A contemporary overview and an analysis of socioeconomic consequences. *Discover Artificial Intelligence, 4*(1), 14. https://doi.org/10.1007/s44163-024–00109-4

Walters, W. H. (2023). The effectiveness of software designed to detect AI-generated writing: A comparison of 16 AI text detectors. *Open Information Science, 7*(1), Article 20220158. https://doi.org/10.1515/opis-2022-0158

Waltzer, T., Pilegard, C., & Heyman, G. D. (2024). Can you spot the bot? Identifying AI-generated writing in college essays. *International Journal for Educational Integrity, 20*(1), 1–18. https://doi.org/10.1007/s40979-024–00158-3

Wan, M. (2023, January 3). Warum ChatGPT nicht das Ende des akademischen Schreibens bedeutet. *digiethics.org.* https://digiethics.org/2023/01/03/warum-chatgpt-nicht-das-ende-des-akademischen-schreibens-bedeutet/

Wang, C., Aguilar, S. J., Bankard, J. S., Bui, E., & Nye, B. (2024). Writing with AI: What college students learned from utilizing ChatGPT for a writing assignment. *Education Sciences, 14*(9), Article 976. https://doi.org/10.3390/educsci14090976

Wang, J., & Fan, W. (2025). The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: Insights from a meta-analysis. *Humanities and Social Sciences Communications, 12*(1), Article 621. https://doi.org/10.1057/s41599-025–04787-y

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity, 19,* Article 26. https://doi.org/10.1007/s40979-023–00146-z

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). *Chain-of-thought prompting elicits reasoning in large language models* (No. arXiv:2201.11903). arXiv. https://doi.org/10.48550/arXiv.2201.11903

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., … Gabriel, I. (2022). Taxonomy of risks posed by language models. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency,* 214–229. https://doi.org/10.1145/3531146.3533088

Wolfram, S. (2023, February 14). *What is ChatGPT doing … and why does it work?* Stephen Wolfram Writings. https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/

World Economic Forum. (2023, September 6). *Can we embrace AI and journey towards a brighter future?* World Economic Forum. https://www.weforum.org/stories/2023/09/can-we-embrace-the-ai-revolution-and-journey-towards-a-brighter-future/

Wu, J., Yang, S., Zhan, R., Yuan, Y., Wong, D. F., & Chao, L. S. (2024). *A survey on LLM-generated text detection: Necessity, methods, and future directions* (No. arXiv:2310.14724). arXiv. https://doi.org/10.48550/arXiv.2310.14724

Xu, X., Tao, C., Shen, T., Xu, C., Xu, H., Long, G., Lou, J., & Ma, S. (2024). *Re-reading improves reasoning in large language models* (No. arXiv:2309.06275). arXiv. https://doi.org/10.48550/arXiv.2309.06275

Xu, Z., Jain, S., & Kankanhalli, M. (2025). *Hallucination is inevitable: An innate limitation of large language models* (No. arXiv:2401.11817). arXiv. https://doi.org/10.48550/arXiv.2401.11817

Yakura, H., Lopez-Lopez, E., Brinkmann, L., Serna, I., Gupta, P., & Rahwan, I. (2024). *Empirical evidence of large language model's influence on human spoken communication* (No. arXiv:2409.01754). arXiv. https://doi.org/10.48550/arXiv.2409.01754

Yan, B., Li, K., Xu, M., Dong, Y., Zhang, Y., Ren, Z., & Cheng, X. (2025). On protecting the data privacy of large language models (LLMs) and LLM agents: A literature review. *High-Confidence Computing, 5*(2), Article 100300. https://doi.org/10.1016/j.hcc.2025.100300

Yan, L., Greiff, S., Teuber, Z., & Gašević, D. (2024). Promises and challenges of generative artificial intelligence for human learning. *Nature Human Behaviour, 8*(10), 1839–1850. https://doi.org/10.1038/s41562-024–02004-5

Yang, K., Li, H., Chu, Y., Lin, Y., Peng, T.-Q., & Liu, H. (2025). *Unpacking political bias in large language models: A cross-model comparison on U.S. politics* (No. arXiv:2412.16746). arXiv. https://doi.org/10.48550/arXiv.2412.16746

Yang, M. (2025, July 12). Elon Musk's AI firm apologizes after chatbot Grok praises Hitler. *The Guardian*. https://www.theguardian.com/us-news/2025/jul/12/elon-musk-grok-antisemitic

Yang, T.-C., Hsu, Y.-C., & Wu, J.-Y. (2025). The effectiveness of ChatGPT in assisting high school students in programming learning: Evidence from a quasi-experimental research. *Interactive Learning Environments, 33*(6), 3726–3743. https://doi.org/10.1080/10494820.2025.2450659

Yin, Z., Wang, H., Horio, K., Kawahara, D., & Sekine, S. (2024). *Should we respect LLMs? A cross-lingual study on the influence of prompt politeness on LLM performance* (No. arXiv:2402.14531). arXiv. https://doi.org/10.48550/arXiv.2402.14531

Zhou, J., Müller, H., Holzinger, A., & Chen, F. (2024). Ethical ChatGPT: Concerns, challenges, and commandments. *Electronics, 13*(17), Article 17. https://doi.org/10.3390/electronics13173417

Zhuo, J., Zhang, S., Fang, X., Duan, H., Lin, D., & Chen, K. (2024). *ProSA: Assessing and understanding the prompt sensitivity of LLMs* (No. arXiv:2410.12405). arXiv. https://doi.org/10.48550/arXiv.2410.12405