»Dann weiß man natürlich nicht immer, ob es stimmt, aber ich vertraue dem«

Reflexionen über und Umgangsweisen mit KI-generierten historischen Erzählungen in Digitalien

Alexandra Krebs

1 Einleitung

Das Zitat im Titel dieses Beitrags stammt aus der aktuell laufenden Studie »De-Constructing History in Digital Space« und ist Teil der Pilotbefragungen von Kindern und Jugendlichen zwischen sieben und 16 Jahren in der Schweiz zu ihrem Verständnis von und ihrem Umgang mit historischen Narrationen im digitalen Raum. An dieser Stelle der Befragung erläutert ein Jugendlicher (14 Jahre) seinen Umgang mit der neusten ChatGPT-Version von 2024 und reflektiert, dass er das Sprachmodell zwar nutzt, um (historische) Informationen z. B. für den Geschichtsunterricht, in der Schweiz integriert im Fächerverbund »Natur Mensch Gesellschaft«, zu erhalten bzw. Erzählungen zu produzieren, diese dann aber nicht weiter überprüft, sondern sie unhinterfragt übernimmt (»ich vertraue dem«).

Im Beitrag steht daher die Frage im Fokus, inwieweit und wie Sprachmodelle historische Erzählungen im digitalen Raum verändern und welche Umgangsweisen mit diesen möglich und notwendig erscheinen. Im ersten Teil werden dafür zunächst die Merkmale, Funktionsweisen und Mechanismen der sogenannten »Large Language Models« (LLM) erläutert und diese zudem in ihrer historischen Entwicklung verortet. Danach gilt es aktuelle Trends und Zukunftsversionen aufzuzeigen sowie vor allem diese kritisch, etwa mit Blick auf »Umweltrassismus« und Diskriminierung sowie Ausschluss von Minderheiten (vgl. Abschnitt 2), zu reflektieren.

Der zweite Teil bezieht diese Überlegungen konkret auf das historische Erzählen im digitalen Raum. Hierfür werden Ausschnitte aus dem beschriebe-

nen Interview exemplarisch vor- und zur Diskussion gestellt, sodass aus geschichtsdidaktischer Perspektive Ableitungen für die Gestaltung historischer Lernprozesse im Umgang mit KI-generierten historischen Erzählungen entwickelt werden können.

2 Stochastische Papageien und das Problem der »Computergläubigkeit«

Die ersten Ideen und Konzepte für unsere heutigen Sprachmodelle wurden bereits 1949 entwickelt. Die frühesten Modelle wurden jedoch erst mehrere Jahre später implementiert. Sie dienten vor allem automatisierter Spracherkennungssoftware, maschineller Übersetzungen oder Textklassifikationen.[1]

1966 präsentierte Joseph Weizenbaum, Professor für Informatik am berühmten MIT, ELIZA. Das Computerprogramm war das erste, welches natürliche Sprache verarbeiten konnte. Hierzu nutzt es allgemeine Methoden zur Analyse von Sätzen und Satzfragmenten, zum Auffinden von sogenannten Schlüsselwörtern in Texten, zum Zusammensetzen von Sätzen aus Fragmenten. Stellt man eine Frage, durchsucht das Programm ein Wörterbuch und verwendet zugleich hinterlegte und vorformulierte Sätze, eine Art Skript als Antwort. Weizenbaum verglich ELIZA daher mit einer Schauspielerin, »die eine Reihe von Techniken beherrschte, aber nichts Eigenes zu sagen hatte.«[2]

ELIZA wird auch als Prototyp moderner Bots wie z.B. ChatGPT bezeichnet. Weizenbaums erste Demonstration seines Produkts lief jedoch nicht so ab, wie er es sich gewünscht hatte. Die Anwesenden waren sehr beeindruckt und verstanden nicht, dass sie mit einer Maschine und nicht mit einem Menschen kommunizierten:

> »Nevertheless, ELIZA created the most remarkable illusion of having understood in the minds of the many people who conversed with it. […] This illusion was especially strong and most tenaciously clung to among people who knew little or nothing about computers. They would often demand to be

1 Vgl. Rosenfeld, Ronald: Two decades of statistical language modeling: Where do we go from here? In: Proceedings of the IEEE 88 (2000), S. 1270–1278.

2 Weizenbaum, Joseph: Computer Power and Human Reason. From Judgement to Calculation. San Francisco 1976, S. 188.

> permitted to converse with the system in private, and would, after conversing with it for a time, insist, in spite of my explanations, that the machine really understood them.«[3]

Der Forscher war schockiert und wurde seitdem zu einem entschiedenen Kritiker, er nannte sich selbst sogar Ketzer der Informatik. In seiner langjährigen Forschertätigkeit sowohl in den USA als auch bis zuletzt in Deutschland warnte er vor den Risiken und Nebenwirkungen einer unreflektierten Nutzung immer größerer und komplexeren Computersysteme:

> »Der Informatiker hat daher die schwerwiegende Verantwortung, die Fehlbarkeit und Begrenztheit der Systeme, die er entwerfen kann, äußerst klarzumachen. Gerade die Wirkungsmöglichkeiten seiner Systeme sollten ihn zögern lassen, bereitwillig seinen Rat zu erteilen, und sollten ihn veranlassen, den Wirkungskreis seiner geplanten Arbeit einzuschränken.«[4]

Seine Appelle, verantwortungsbewusst mit derartigen neuen Entwicklungen umzugehen und zugleich vor allem mögliche gesellschaftliche Gefahren zu reflektieren und sich ggf. auch gegen eine neue Idee zu entscheiden, wurden zwar bis zu seinem Tod 2008 vielfach, vor allem in der Wissenschaft rezipiert, scheinen jedoch weitgehend verhallt zu sein, wie im Folgenden noch zu zeigen ist.

Zuvor stellt sich die Frage, wie sich die Sprachmodelle von ELIZA bis zu neusten ChatGPT Version von OpenAI weiterentwickelt haben. Zunächst einmal unterscheiden sich die neu entwickelten Programme, also v.a. die Generationen seit 2018, von der ELIZA-Generation darin, dass sie keine Wörterbücher durchsuchen und auch nicht auf vorformulierte Sätze zurückgreifen. Es handelt sich meist um Modelle, welche im Dialog mit den Nutzenden Inhalte mithilfe von KI generieren (engl. »artificial intelligence generated content model«).[5] Das heißt, die Modelle »erlernen« eigenständig mit Sprache umzugehen, indem sie mit großen Textmengen »trainieren«. Sie basieren auf Mustererkennung und Übergangswahrscheinlichkeiten. Sie können also die Abfolge von »strings« (Buchstabenfolgen) vorhersagen. Oder

3 Ebd., S. 189.

4 Weizenbaum, Joseph, 1972: Alptraum Computer, https://www.zeit.de/1972/03/alptraum-computer, aufgerufen am 14.02.2024.

5 Vgl. Ouyang, Long u.a., 2022: Training language models to follow instructions with human feedback, https://arxiv.org/abs/2203.02155, aufgerufen am 14.02.2024, S. 1122.

anders ausgedrückt: LLM sagen die Wahrscheinlichkeit, dass ein bestimmter »token« (Zeichen, Wort oder Zeichenfolge) entweder auf einen vorhergehenden oder einen umgebenden Kontext folgt.[6] Sehr vereinfacht ausgedrückt, lernt das Sprachmodell

> »aufgrund einer mathematisch-statistischen Analyse, welches Zeichen oder welches Wort in einem Text mit großer Wahrscheinlichkeit als nächstes folgt. […] Sind mehrere Wörter ähnlich wahrscheinlich, wird in der Regel mit einer gewissen Zufälligkeit ein Wort ausgewählt.«[7]

Daher entstehen auch immer unterschiedliche Ergebnisse sowie inhaltliche Fehler.[8] SoekiaGPT, ein didaktisches Sprachmodell, macht diesen Prozess – wenn auch sehr vereinfacht und daher nicht gänzlich mit den aktuellen LLM vergleichbar – deutlich: Basieren die Trainingsdaten z.B. auf Märchenerzählungen, würde auf den Satzanfang »Es war einmal ein« mit einer Häufigkeit von 0.10 das Wort »Müller«, mit 0.09 »König«, mit 0.08 »Bauer«, mit 0.06 »fröhlicher«, mit 0.03 »grüner« usw. folgen. Weshalb das Modell den Satz »Es war einmal ein Müller« vervollständigt und danach wiederum das nächste Wort sucht.[9]

Daran wird ersichtlich, dass die Modelle – sei es nun ChatGPT oder ELIZA – kein Bewusstsein und kein Sprachverständnis besitzen, sondern allein Statistik die Wortfolge der Sätze bestimmt und Wahrscheinlichkeitsrechnung das einzige ist, das sie beherrschen:

> »However, no actual language understanding is taking place in LM-driven approaches […] languages are systems of signs, i.e. pairings of form and

6 Bender, Emily M; Gebru, Timnit; McMillan-Major, Angelina; Shmitchell, Shmargaret: On the Dangers of Stochastic Parrots. In: Conference on Fairness, Accountability, and Transparency (FAccT '21), March 3–10, 2021, Virtual Event, Canada (2021), S. 610–623, hier S. 611.

7 Hielscher, Michael: SoekiaGPT – ein didaktisches Sprachmodell. In: Informatische Bildung in Schulen 1 (2023), H. 1, S. 1–11, hier S. 3.

8 Da die Modelle nicht auf Informationen (wie etwa bei einer Suchmaschine) zurückgreifen, sondern auf Mustererkennung und Übergangswahrscheinlichkeiten (s. Erklärung oben im Text), finden sich in den Texten z.B. Bücher und Literatur sowie Forschungsaufsätze, welche nicht existieren sowie grundsätzlich Fehler in vielerlei Bereichen, v.a. wenn sie nicht zum »Allgemeinwissen« gehören und daher in den Trainingsdaten weniger umfangreich vorhanden sind.

9 Hielscher, SoekiaGPT – ein didaktisches Sprachmodell, S. 3.

> meaning. But the training data for LMs is only form; they do not have access to meaning.«[10]

Um dies zu verstehen, hilft auch die folgende Erläuterung von Murray Shanahan: Geben wir z.B. einem LLM den Prompt »Die erste Person auf dem Mond war«, erwarten wir als Antwort «Neil Armstrong«. Tatsächlich fragen wir hier aber nicht, wer die erste Person auf dem Mond war, sondern wir fragen: »Welche Wörter folgen angesichts der statistischen Verteilung von Wörtern im riesigen öffentlichen Textkorpus am ehesten auf die Sequenz »Die erste Person auf dem Mond war«? Am ehesten folgt hierauf nun mal »Neil Armstrong«.[11] Sprachmodelle lassen sich also gut mit dem Bild »stochastischer Papageien« vergleichen.[12] Sie führen willkürlich Sequenzen sprachlicher Formen zusammen, welche die Modelle in den umfangreichen Trainingsdaten beobachtet haben, und zwar auf der Grundlage probabilistischer Informationen darüber, wie sie kombiniert werden, aber ohne jeglichen Bezug zur Bedeutung.[13]

Mit der Größe der Trainingsdaten wachsen jedoch – zumindest scheinbar – die Fähigkeiten der Modelle, d.h. ihre »performance«[14] verbessert sich stetig. ChatGPT-4 produziert mittlerweile so überzeugende Texte, dass ohne Frage leicht die Illusion entsteht, mit einer menschlich-denkenden Kreatur zu interagieren[15] oder sich zumindest zu fragen, ob nicht doch mehr dahinterstecken könnte als bloße Wahrscheinlichkeitsberechnungen. Genauso wie Weizenbaum 1966 stehen wir heute also wieder – und vielleicht sogar noch viel mehr – vor der Herausforderung, KI-getriebene Programme nicht mit menschlichen Fähigkeiten und Verhaltensweisen gleichzusetzen:

> »[…] the point is that such systems are simultaneously so very different from humans in their construction yet (often but not always) so human-like in their behavior, that we need to pay careful attention to how they work be-

10 Bender u.a., On the Dangers of Stochastic Parrots, S. 615.

11 Shanahan, Murray: Talking about Large Language Models. In: Commun. ACM 67 (2024), H. 2, S. 68–79, hier S. 70.

12 Bender u.a., On the Dangers of Stochastic Parrots.

13 Ebd., S. 617.

14 Dies meint das Abschneiden der Modelle in bestimmten, standardisierten Tests. Vgl. u.a. OpenAI u.a., 2023: GPT-4 Technical Report, https://arxiv.org/pdf/2303.08774, aufgerufen am 14.02.2024.

15 Vgl. Shanahan: Talking about Large Language Models, S. 79.

> fore we speak of them in language suggestive of human capabilities and patterns of behavior.«[16]

Wichtig ist Anthropomorphismen, also Vermenschlichungen, bei der Beschreibung von und im Umgang mit den Programmen zu vermeiden. Z.B. »spricht« die KI nicht mit uns oder »hat sich etwas bei ihrer Aussage gedacht«. Oftmals personifizieren wir die Modelle, wenn wir etwa sagen »Jetzt ist er ins Stocken geraten« oder »Es hat mir eine gute Antwort gegeben.«[17] Vielmehr sollten wir versuchen auch in unserem Sprechen über die Computersysteme zu reflektieren, dass diese kein »Selbst« besitzen, keine menschlichen Wesen sind.

Dies ist sicherlich nicht immer einfach, denn seit der Entwicklung der ersten Version von ChatGPT-1 2018 hat jedes weitere Update den Funktionsumfang erweitert, etwa für verschiedene Sprachverstehens-Aufgaben und Generierungen, wie z.B. mehrsprachige maschinelle Übersetzung, Code-Debugging, Schreiben von Geschichten, Eingestehen von Fehlern und sogar Zurückweisen von unangemessenen Anfragen. Besonders seit 2023, mit der Veröffentlichung von ChatGPT-4 durch OpenAI hat sich der Funktionsumfang des LLM weiter vergrößert und zugleich verbessert. Die Zahl der Nutzenden überschritt daher auch schnell die 100 Millionen-Grenze. Sie können nun nicht mehr nur Text, sondern Text und Bilder zugleich eingeben sowie multimodale Aufgaben stellen, etwa Bildbeschriftungen, Diagrammerstellung und Zusammenfassung wissenschaftlicher Publikationen.[18]

Neben den LLM wurden jedoch vor allem auch seit 2022 vielfältige weitere Produkte aus der Familie der »artificial intelligence generated content models« entwickelt, wie z.B. DALL-E-2 (ebenso von OpenAI), mit welchem sich KI-generierte Bilder von teils täuschend echter Qualität erzeugen lassen. Mit dem

16 Ebd., S. 71.

17 Walter, Yoshija, 2023: Die Vermenschlichung der künstlichen Intelligenz. Die grosse sozio-psychologische Kritik und das KI-Bewusstsein, https://www.kalaidos-fh.ch/de-CH/Blog/Posts/2023/11/Digitalisierung-1121-Vermenschlichung-kuenstliche-Intelligenz., aufgerufen am 14.02.2024.

18 Vgl. Ouyang u.a., Training language models to follow instructions with human feedback, S. 1122.

Meta-Produkt »Make-A-Video« lässt sich zudem Text in Video verwandeln[19] – um nur eine kleine Auswahl zu erwähnen.[20]

Haben wir es also mit einer kulturellen Revolution zu tun, einem Meilenstein der Informatik und Technikentwicklung? Betrachtet man allein die neuen Möglichkeiten und Funktionen der Programme, ihr Abschneiden in verschiedenen Tests[21] kann man dies sicherlich bejahen. Wir sollten uns jedoch auch fragen, welche Risiken und Folgen damit verbunden sind und wie wir folglich mit ihnen umgehen sollten.

3 »There is no data like more data«

Zunächst ist es wichtig zu beschreiben, wie und durch welche Faktoren sich die neuen Programme überhaupt entwickeln konnten. Der AI Index Report 2024 der Stanford University in den USA gibt hierüber Aufschluss. Der Bericht sammelt und analysiert Daten zum Thema KI mit dem Ziel, diese v.a. für einen möglichst breitgefächerten Adressatenkreis, wie z.B. Journalist*innen, Politiker*innen sowie Forschende aus verschiedenen Disziplinen, zugänglich und verständlich zu machen.[22]

Aus diesem Bericht stammend sind in Abbildung 1 die 2023 bekanntesten Machine Learning Modelle nach Ländern aufgelistet. Die USA liegen mit großem Abstand vorne (61 Modelle), gefolgt von China (15 Modelle), Frankreich (8), Deutschland (5) und Kanada (4) auf den ersten fünf Plätzen. Auffallend ist hierbei, dass es sich bei diesen ausschließlich um Länder des Globalen Nordens handelt und der Globale Süden nicht vertreten ist.[23]

19 Ebd.

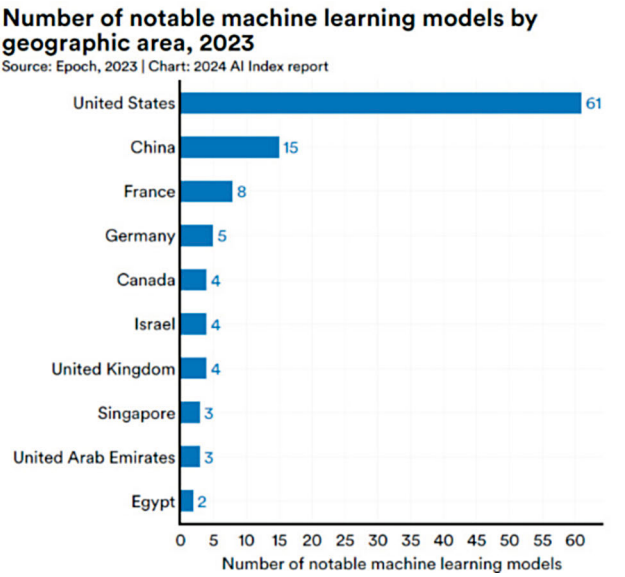
20 Eine Übersicht findet sich bei Maslej, Nestor u.a.: The AI Index 2024 Annual Report. AI Index Steering Committee, Institute for Human-Centered AI. Stanford 2024, S. 78–80.

21 Vgl. hierzu OpenAI u.a., GPT-4 Technical Report sowie den AI Index Report 2024: »Over the years, AI has surpassed human baselines on a handful of benchmarks, such as image classification in 2015, basic reading comprehension in 2017, visual reasoning in 2020, and natural language inference in 2021.« Maslej u.a., The AI Index 2024 Annual Report, S. 81.

22 Vgl. Maslej u.a., The AI Index 2024 Annual Report, S. 2.

23 Ebd., S. 47f.

Abbildung 1: Anzahl bekannter Machine Learning Modelle nach Ländern (ebd. S. 47).



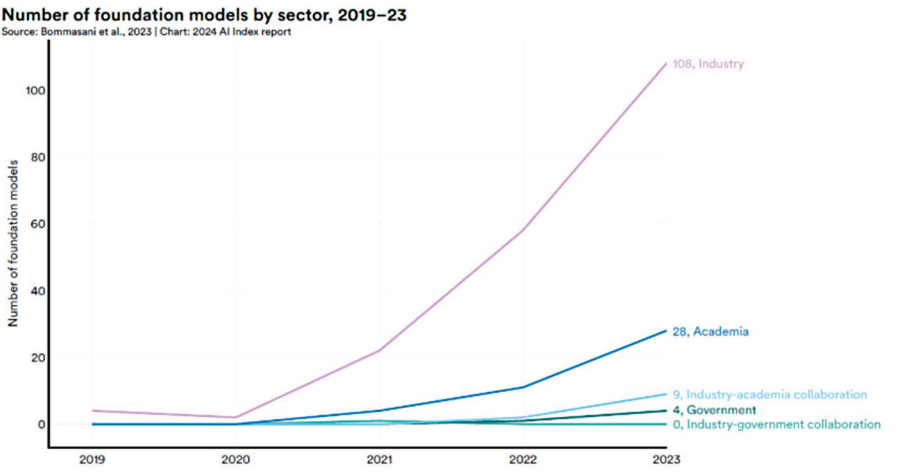
Ebenso beachtenswert ist, dass die meisten der Modelle (72,5 %) im industriellen Sektor entwickelt werden und nur ein wesentlich kleinerer Teil (18 %) im akademischen, also an Universitäten und Forschungseinrichtungen (vgl. Abbildung 2).

Zugleich zeigt sich ein enormer Anstieg der Trainingskosten für die immer umfangreicheren und komplexeren Modelle (vgl. Abbildung 3 und 4). Die in Abbildung 3 und 4 dargestellten Zahlen basieren auf Schätzungen. Oftmals legen KI-Firmen und Startups ihre Kosten nämlich nicht offen, es ist jedoch davon auszugehen, dass diese mittlerweile Millionen von Dollar bei weitem überschreiten. Das multidisziplinäre US-Forschungszentrum »Epoch AI«[24] mit Sitz in San Francisco analysierte u.a. die Trainingsdauer sowie die Art, Menge und Nutzungsrate der Trainingshardware anhand von Informationen aus Veröffentlichungen, Pressemitteilungen oder technischen Berichten

24 Epoch AI ist ein Forschungsinstitut, das u.a. Trends und Fragen in Bezug auf die Entwicklung und den Gebrauch von KI betreffen. O.A.: Epoch AI, https://epochai.org/, aufgerufen am 14.02.2024.

zu den Modellen.[25] Deutlich wird ein exponentielles Wachstum der Kosten, die also sprichwörtlich durch die Decke gehen: Lagen die Trainingskosten 2017 für das Transformer-Modell (dieses war das erste mit der neuen Modell-Architektur der aktuellen LLM) noch bei überschaubaren $ 900, summierten sich z.B. für RoBERTa Large 2019 bereits $ 160 000. Für 2023 wurden die Kosten für ChatGPT-4 sowie Googles Gemini Ultra auf 78 Millionen bzw. 191 Millionen US-Dollar geschätzt (vgl. Abbildung 3 und 4).[26]

Abbildung 2: Entwicklung der Anzahl von Grundlagenmodellen (»foundation models«) nach Sektoren (ebd. S. 58).



In nur wenigen Jahren haben sich die Kosten um mehr als das 200 000-fache vergrößert. Die Gründe hierfür sind, dass immer größere Trainingsdaten und umfangreichere Parameter zum Trainieren der Modelle genutzt werden. 2018 beliefen sich z.B. die Trainingsdaten für ChatGPT-1 noch auf 5 GB und die Parameter auf 117 Millionen, 2019 für ChatGPT-2 waren es bereits 40 GB, und 1,5 Billionen Parameter, 2020 für ChatGPT-3 45 TB und 175 Billionen Parameter.[27] Die neuesten Modelle, wie ChatGPT-4 und Gemini, benötigen Schät-

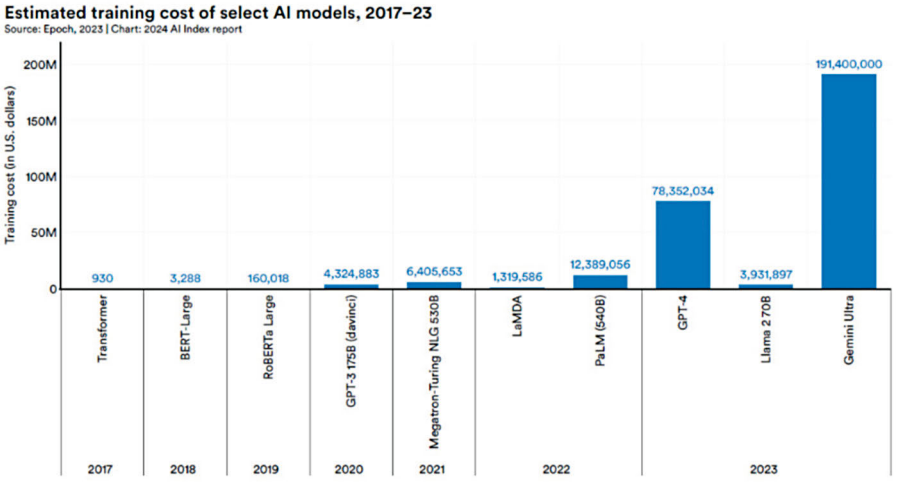
25 Vgl. Maslej u.a., The AI Index 2024 Annual Report, S. 63.

26 Ebd., S. 61.

27 Vgl. Ouyang u.a., Training language models to follow instructions with human feedback, S. 1123.

zungen zufolge noch um ein Vielfaches größere Datensätze. Nach der Devise »there is no data like more data« wird sich diese Entwicklung immer weiter fortsetzen, solange eine wachsende Größe der Trainingsdaten mit einer besseren »performance« der Modelle in den verschiedenen Aufgabenbereichen korreliert.[28]

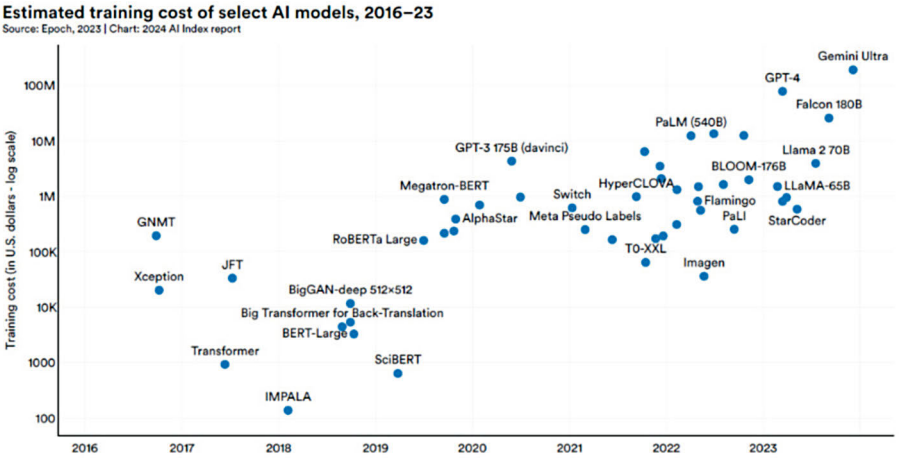
Abbildung 3: Geschätzte Summe der Trainingskosten ausgewählter KI-Modelle (2016–2023) (Maslej u.a. 2024, S. 64).



Die gestiegenen Kosten haben Einfluss darauf, wer bzw. welche Institutionen es sich überhaupt leisten können, neue Modelle zu entwickeln. Dies erklärt sowohl die geographische Verteilung der Modelle (überwiegend im Globalen Norden, vgl. Abbildung 1) als auch der Sektoren (überwiegend marktwirtschaftlich- und gewinnorientierte Industrie, vgl. Abbildung 2).

28 Bender u.a., On the Dangers of Stochastic Parrots, S. 611.

Abbildung 4: Geschätzte Summe der Trainingskosten ausgewählter KI-Modelle (2017–2023) auf einer logarithmischen Skala (ebd.).



Welche Konsequenzen ergeben sich nun aber aus diesen Entwicklungen und Möglichkeiten der neuen LLM?

Nach Bender u.a. lassen sich verschiedene Konsequenzen hieran aufzeigen, welche von der Forschungsgruppe als globale Risiken eingestuft werden. Zunächst verursacht das Training der immer größer werdenden Modelle auch immer immensere Kosten, welche sich auch auf Umwelt und Klimawandel auswirken. Die Kosten der klimatischen Veränderungen tragen jedoch vor allem jene Menschen des Globalen Südens, die bereits benachteiligt sind und zugleich von den Vorteilen der neuen technischen Entwicklungen weniger bis gar nicht profitieren, denn oftmals haben sie keinen Zugriff auf die Sprachmodelle, da in diesen ihre Sprachen meist nicht enthalten sind (wie z.B. Dhivehi oder Sudanesisch-Arabisch).[29] Dies lässt sich als eine Form des »Umweltrassismus« beschreiben, d.h. ohnehin bereits marginalisierte Gruppen sind besonders stark von den negativen Folgen und Auswirkungen des Klimawandels betroffen.[30]

29 Ebd., S. 612f.

30 Vgl. hierzu u.a. das Grundlagenwerk von Westra, Laura; Lawson, Bill E.: Faces of environmental racism. Confronting issues of global justice. 2. Aufl. Lanham 2001 (Studies in social, political, and legal philosophy).

Ein zweites Risiko hängt mit den undurchsichtigen Trainingsdaten der Sprachmodelle zusammen. Damit die Modelle »lernen« können, Sprache zu imitieren, benötigen sie natürliche, d.h. von Menschen verfasste, Texte in möglichst vielfältigen Formen und mit möglichst umfangreichen Inhalten. Zu Beginn, also z.B. für ChatGPT-1, wurden vor allem frei zugängliche Texte aus der Wikipedia sowie dem BookCorpus-Datensatz, der sich aus etwa 7 000 selbstveröffentlichten Büchern auf der E-Book Webseite »Smashwords« zusammensetzt, genutzt. Für die weiteren Versionen und Trainingsdatensätze kamen z.B. Webtext[31] sowie Common Crawl[32] hinzu.[33] Man könnte vermuten, dass solche Daten eine Art weltweite Vielfalt aus unterschiedlichsten Perspektiven darstellen, da das Internet grundsätzlich allen offen steht. Dies ist jedoch ein Trugschluss. Nicht alle Menschen weltweit haben Zugriff auf das Internet und können dort Inhalte verfassen. Dies ist vor allem der jüngeren Weltbevölkerung als auch jener aus westlichen Industriestaaten vorbehalten. Zudem haben Untersuchungen gezeigt, dass z.B. Wikipedia-Artikel hauptsächlich von Männern und nur zwischen 8,8-15 % der Texte von Frauen verfasst wurden, und dass die Daten, wie jene der Plattform Reddit, z.B. in den USA zu 67 % von Männern im Alter zwischen 18 und 29 stammen.[34] Es handelt sich also keinesfalls um repräsentative Daten, welche gesellschaftliche Diversität und Pluralität widerspiegeln. Vielmehr sind diese nicht kurartierten Daten geprägt von einer dominanten, hegemonialen Sichtweise und aufgrund ihres Zuschnittes und fehlender Filter voll von Stereotypen, Rassismus, Antisemitismus, Sexismus, Misogynie usw.:

---

31 WebText ist ein internes OpenAI-Korpus, das durch Scraping von Webseiten erstellt wurde. Die Autor*innen sammelten alle ausgehenden Links von Reddit, die andere Nutzer*innen interessant, lehrreich oder einfach nur lustig fanden. WebText enthält die Textteilmenge dieser 45 Millionen Links. Sie besteht aus über 8 Millionen Dokumenten mit einer Gesamtmenge von 40 GB Text (o.A.: WebText, https://paperswithcode.com/dataset/webtext, aufgerufen am 14.02.2024).

32 Common Crawl wurde ebenso durch Webscraping seit 2007 zusammengestellt. Der Korpus ist frei und offen zugänglich und kann z.B. für Forschung und Entwicklung genutzt werden. Er umfasst über 250 Billionen Webeseiten (o.A.: Common Crawl, https://commoncrawl.org/, aufgerufen am 14.02.2024).

33 Vgl. Ouyang u.a., Training language models to follow instructions with human feedback, S. 1123.

34 Vgl. Barera, Michael, 2020: Mind the Gap: Addressing Structural Equity and Inclusion on Wikipedia, http://hdl.handle.net/10106/29572, aufgerufen am 14.02.2024.

> »In the case of US and UK English, this means that white supremacist and misogynistic, ageist etc. views are overrepresented in the training data, not only exceeding their prevalence in the general population but also setting up models trained on these datasets to further amplify biases and harms.«[35]

Nutzen wir also Sprachmodelle, welche überrepräsentativ Texte mit inhärentem Bias imitieren, reproduzieren wir in den neu erstellen Texten diese wiederum und verstärken ihre Reichweite und die negativen Folgen dadurch. Vor Augen geführt wird dies besonders bei Text-zu-Bild-Modellen, wie etwa Stable Diffusion.[36] Die KI-generierten Bilder spiegeln nicht nur z.B. rassistische und geschlechterspezifische Stereotypen wider, sie überzeichnen sie sogar um ein Vielfaches, wie verschiedene Testungen in Bezug auf Berufsgruppen gezeigt haben:

> »The world according to Stable Diffusion is run by White male CEOs. Women are rarely doctors, lawyers or judges. Men with dark skin commit crimes, while women with dark skin flip burgers.«[37]

Die KI-genierten Bilder waren im untersuchten Datensatz also sogar noch problematischer als unsere Realität.[38] Einige der führenden KI-Unternehmen sind sich dieser Problematik mittlerweile auch bewusst geworden und versuchen Gegenmaßnahmen zu ergreifen, man nennt dies »Alignment«, also die

35 Bender u.a., On the Dangers of Stochastic Parrots, S. 613.

36 Das Text-zu-Bild-Modell wurde von einer Forschungsgruppe der LMU München entwickelt. Auf der Webseite findet sich folgender Beschreibungstext: »Stable Diffusion ist ein latentes Text-zu-Bild-Difusionsmodell, das in der Lage ist, fotorealistische Bilder aus jeglicher Texteingabe zu generieren, fördert die autonome Freiheit, um unglaubliche Bilder zu produzieren, und ermöglicht es Milliarden von Menschen, beeindruckendes Kunstwerk in Sekunden zu erstellen.« (O.A.: Stable Diffusion, https://stablediffusionweb.com/de, aufgerufen am 14.02.2024).

37 Nicoletti, Leonardo; Bass, Dina, 2023: Humans Are Biased. Generative AI is even worse, https://www.bloomberg.com/graphics/2023-generative-ai-bias/, aufgerufen am 14.02.2024.

38 Am Arbeitsbereich Didaktik der Geschichte der Universität Marburg arbeitet Thorsten Neischwander an einer Dissertation, welche die Möglichkeiten historischer Imaginationskompetenzen von Lernenden mit Hilfe bildgenerierender KI (MidJourney) erforscht. Für empirische Forschungen im Bereich historischen Lernens eröffnen sich dank KI neue Möglichkeiten.

Fehler der KI ausbessern, »sie auf Linie zu bringen«.[39] Zuletzt fiel dies beim Modell Gemini von Google auf. Den Prompt »Bilder von Päpsten zu generieren« beantwortete das Modell mit Darstellungen einer Schwarzen Frau und einem Schwarzen Mann. Google hatte wohl einen bestimmten Automatismus eingebaut:

> »Im Hintergrund wurde offenbar jede der Eingaben, sogenannte Prompts, um einen unsichtbaren Hinweis ergänzt. […] Die Aufforderung »Zeige mir einen Papst« wurde demnach an das System weitergegeben mit der Aufforderung, »explizit verschiedene Geschlechter und Ethnien« einzubeziehen.«[40]

Ähnliches gilt wohl auch für die von Gemini entworfenen Wehrmachtssoldaten mit nicht weißer Hautfarbe. Hieran wird deutlich, wie schwierig es ist, die Verzerrungen und Diskriminierungen der Datensätze nachträglich auszubessern. Zudem haben weitere Untersuchungen gezeigt, dass sie oftmals »im Hintergrund« in den Modellen, trotz der Ausbesserungsversuche, enthalten bleiben.[41]

Wenn wir immer nur auf die Fehler der Modelle reagieren, laufen wir womöglich Gefahr, unsere Handlungsfähigkeiten zu verlieren: »Feeding AI systems on the world's beauty, ugliness, and cruelty, but expecting it to reflect only the beauty is a fantasy.«[42] Dieser Hoffnung und Fantasie, dass die Modelle selbst ihre Unzulänglichkeiten überwinden, sollten wir uns also nicht allzu leichtfertig hingeben, sondern auch die Frage stellen, ob wir tatsächlich immer größere Sprachmodelle benötigen, wofür wir sie eigentlich einsetzen sollten und wofür nicht, ob es sinnvoll ist, weitere Forschung in diese Richtung zu betreiben und immer mehr Geld in diese Entwicklung zu investieren. Auch wenn uns die Modelle in den Tests ein vermeintliches Sprachverständnis mittlerweile immer besser vortäuschen, bleiben sie »stochastische Papageien«,[43] die lediglich Sprache imitieren:

---

39 Lindern, Jakob von; Wolfangel, Eva, 2024: Ist das der Papst?, https://www.zeit.de/2024/13/diversitaet-google-ki-gemini-bild-generator-papst, aufgerufen am 14.02.2024.

40 Ebd.

41 Hubinger, Evan u.a., 2024: Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training, http://arxiv.org/pdf/2401.05566v3, aufgerufen am 14.02.2024.

42 Bender u.a., On the Dangers of Stochastic Parrots, S. 615.

43 Ebd.

> »If a large LM, endowed with hundreds of billions of parameters and trained on a very large dataset, can manipulate linguistic form well enough to cheat its way through tests meant to require language understanding, have we learned anything of value about how to build machine language understanding or have we been led down the garden path?«[44]

Erst kürzlich ist hierzu eine neue Publikation erschienen, welche bereits im Titel diese Problematik anschaulich auf den Punkt bringt:

»Alice in Wonderland: Simple Tasks Showing Complete Reasoning Breakdown in State-Of-the-Art Large Language Models«.[45] Die Forschungsgruppe deckt auf, dass die zuvor beschriebenen hohen Punktzahlen der neusten LLM in den standardisierten Testungen keinesfalls natürliches Sprachverstehen demonstrieren, da sie bei simplen, logischen Schlussfolgerungen katastrophal scheitern:[46]

> »The breakdown is dramatic, as models also express strong overconfidence in their wrong solutions, while providing often non-sensical »reasoning«-like explanations akin to confabulations to justify and backup the validity of their clearly failed responses, making them sound plausible.«[47]

Neben den bereits beschriebenen Risiken hat dies zur Folge, dass Menschen immer mehr den Modellen vertrauen und zutrauen sowie deren Antworten immer weniger hinterfragen.[48] Das eingangs zitierte Fallbeispiel aus den Inter-

44 Ebd., S. 616.

45 Nezhurina, Marianna; Cipolina-Kun, Lucia; Cherti, Mehdi; Jitsev, Jenia: Alice in Wonderland, 2024: Simple Tasks Showing Complete Reasoning Breakdown in State-Of-the-Art Large Language Models, https://arxiv.org/pdf/2406.02061, aufgerufen am 04.09.2024.

46 Die genutzte Frage im Test lautete: »Alice has 4 brothers and she also has 1 sisters. How many sisters does Alice's brother have?«, ebd., S. 3.

47 Ebd., S. 1.

48 Dies birgt auch die Gefahr eines bewussten Missbrauchs der Technologie, z.B. durch extremistische Gruppen: » […] GPT-3 could be used to generate text in the persona of a conspiracy theorist, which in turn could be used to populate extremist recruitment message boards. This would give such groups a cheap way to boost recruitment by making human targets feel like they were among many like-minded people. If the LMs are deployed in this way to recruit more people to extremist causes, then harms, in the first instance, befall the people so recruited and (likely more severely) to others as a result of violence carried out by the extremists.« Bender u.a., On the Dangers of Stochastic Parrots, S. 617.

views mit Lernenden in der Schweiz kann den Umgang mit LLM im Kontext des historischen Denkens im nun folgenden Teil des Beitrags zumindest exemplarisch verdeutlichen.

4 Reflexionen über und Umgangsweisen mit historischen Erzählungen im digitalen Raum

Das Projekt »De-Constructing History in Digital Space« untersucht u.a. im Kontext der zuvor beschriebenen neueren technischen Entwicklungen, den Umgang von Kindern und Jugendlichen zwischen sieben und 16 Jahren in der Schweiz mit historischen Narrationen im digitalen Raum.

Bereits in zahlreichen, teils auch schon älteren empirischen Studien wurde deutlich, dass Lernende historische Darstellungen im Netz oftmals unhinterfragt übernehmen. Jan Hodel zeigte z.B., dass Schüler*innen in ihren Referaten Narrationen lediglich reproduzieren, d.h. sie entnehmen Informationen etwa aus Wikipedia-Seiten, wählen verschiedene Fragmente aus und unterziehen diese nicht »einer expliziten Überprüfung nach fachlichen Kriterien der empirischen Triftigkeit.«[49] Zu ähnlichen Ergebnissen gelangte Sam Wineburg in den USA. Er untersuchte, nach welchen Kriterien Lernende Informationen aus dem Internet auswählen:

> »Ausschlaggebend […] war nicht die Vertrauenswürdigkeit der geschichtlichen Informationen, ob die Interpretation schlüssig war […]. Entscheidend war, ob die Informationen sich in die iBook-Vorlage von Apple einfügen ließen.«[50]

Junge Lernende (ebenso wie auch Erwachsene) scheinen also Schwierigkeiten zu haben, reflektiert mit Darstellungen umzugehen. Dies wurde auch in der Studie im Kontext historischer Lernprozesse auf der Plattform »App in die Geschichte« deutlich: Zwar machen ein Großteil der Lernenden die Nutzung von

49 Hodel, Jan: Verkürzen und Verknüpfen. Geschichte als Netz narrativer Fragmente: Wie Jugendliche digitale Netzmedien für die Erstellung von Referaten im Geschichtsunterricht verwenden. Bern 2013 (Geschichtsdidaktik heute, Bd. 5), S. 327.

50 Wineburg, Sam: Warum historische Kompetenzen für die Auswertung von digitalen Quellen nicht ausreichend sind. In: Sebastian Barsch, Andreas Lutter, Christian Meyer-Heidemann (Hg.): Fake und Filter. Historisches und politisches Lernen in Zeiten der Digitalität. Frankfurt a.M. 2019 (Wochenschau Wissenschaft), S. 105–120, hier S. 105.

Darstellungen in ihren Narrationen in Form von Zitaten und Verweisen transparent, allerdings erkennt nur ein kleiner Teil von ihnen überhaupt die Kontroversität historischer Darstellungen und reflektiert bzw. analysiert diese explizit.[51]

Der Umgang mit Geschichten sowie insbesondere deren De-Konstruktion, also eine kritisch-reflexive Analyse,[52] stellen für Lernende eine Herausforderung dar und sollten daher stärker in den Fokus von Lernprozessen, vor allem auch im digitalen Raum, gerückt werden. Dort prägen nämlich auch Formen von Desinformation, Propaganda oder sogar Verschwörungserzählungen historische Narrative, insbesondere in den Social Media. Vor allem rechtsextremistische Gruppen nutzen derartige Praktiken und Phänomene im Netz. Sie propagieren oftmals xenophobe und antisemitische Narrationen, »Sozialdarwinismus« und verharmlosen Nationalsozialismus und Holocaust.[53] Gegenwärtige Neuerungen, wie die Entwicklung der LLM, können dies weiter verstärken.[54]

Derlei Herausforderungen im digitalen Raum sind längst Teil der Lebens- und Alltagswelt von Jugendlichen. Sie müssen sich zu ihnen verhalten und mit ihnen umgehen (lernen). Das Projekt »De-Constructing History in Digital Space« führt hierfür verschiedene Disziplinen und Konzepte zusammen, v.a. der Geschichtsdidaktik und Geschichtswissenschaft, ebenso wie der Politikdidaktik und Medienpädagogik. Daran anschließend erforscht das Projekt empirisch, wie Jugendliche verschiedener Altersstufen in der Schweiz kontroverse historische Narrationen im digitalen Raum verhandeln und wie sie mit diesen umgehen. Hierzu werden zunächst Interviews durchgeführt (Pilotphase). Danach untersuchen Schüler*innen verschiedener Schulen sowie Jahrgangsstufen (Jg. 6, 9 und 11) in Tandems kontroverse Narrationen. Dabei werden sowohl Audiomitschnitte als auch Bildschirmaufzeichnungen

51 Vgl. Krebs, Alexandra: Geschichten im digitalen Raum. Historisches Lernen in der »App in die Geschichte«. Berlin 2024 (Medien der Geschichte, Bd. 7), S. 285.

52 Vgl. Schreiber, Waltraud: Kompetenzbereich historische Methodenkompetenz. In: Andreas Körber, Waltraud Schreiber, Alexander Schöner (Hg.): Kompetenzen historischen Denkens. Ein Strukturmodell als Beitrag zur Kompetenzorientierung in der Geschichtsdidaktik. Neuried 2007 (Kompetenzen: Grundlagen – Entwicklungen – Förderung, Bd. 2), S. 194–235, hier S. 224–230.

53 Vgl. Schwarz, Karolin: Hasskrieger. Der neue globale Rechtsextremismus. Sonderausgabe für die Bundeszentrale für politische Bildung. Bonn 2020 (Schriftenreihe/ Bundeszentrale für Politische Bildung, Band 10545).

54 Vgl. Anm. 403.

sowie Logfiles erhoben. Die Daten werden in einem Mixed-Method Design mithilfe von Pfadanalysen, Clustering und qualitativer Inhaltsanalyse ausgewertet. Ziel ist es hieraus Ableitungen zu folgern, um historische Lernprozesse und Lernangebote zu gestalten. Im Folgenden wird nun ein Fallbeispiel aus den Interviews der Studie in Bezug auf den Umgang mit Sprachmodellen vorgestellt.

Da beim Umgang mit Sprachmodellen und historischen Erzählungen auch das Geschichtsverständnis eine Rolle spielt, ist es zunächst interessant, wie sich die Jugendlichen dazu äußern, was »Geschichte« eigentlich ist. Im ausgewählten Interview versteht der Jugendliche (14 Jahre) darunter vor allem »[…] Vergangenheit, denke ich, was in der Vergangenheit alles passiert ist.« (P3_14_m, Pos. 39) und betont, dass ihm davon derzeit viel im digitalen Raum begegnet:

> »Ja, es gibt einfach ganz viele Videos auf Social Media, auf Tiktok vor allem oder sicher am meisten, dort siehst du einfach alles gefühlt, was es gibt, so Vergangenheit betreffend« (P3_14_m, Pos. 51).

Deutlich wird vor allem, dass ein solches Geschichtsverständnis als positivistisch eingestuft werden kann. Geschichte sei also die Vergangenheit, das was »passiert ist«. Dieses Geschichtsverständnis unterscheidet sich grundlegend vom narrativ-konstruktivistischen Verständnis, wonach Geschichte stets eine gegenwärtige Konstruktion, also Erzählung, über Vergangenes ist, die sich je nach Perspektive, Standort, Intention der Erzählenden usw. unterscheiden kann. Daher finden sich meist auch plurale Erzählungen, welche sich durchaus auch widersprechen können.[55] Seit dem »narrative turn«[56] der Geschichtswis-

55 Wichtig ist dabei jedoch, dass es nicht um beliebige Geschichten geht, also nicht jede Erzählung gleich zustimmungsfähig ist, sondern um solche, die überzeugender – nach Jörn Rüsen triftiger – als andere sind, nämlich in empirischer, narrativer sowie normativer Hinsicht. Vgl. Krebs, Geschichten im digitalen Raum, S. 39f.

56 Im Kontext des »narrative turn« der Geschichts- und Erkenntnistheorie in den 1980er Jahren entwickelten Vertreter*innen einer sogenannten narrativen Historik einen erweiterten Erzählbegriff, indem sie postulierten, dass »Erzählen und diskursive Vernunfttätigkeit in den geistigen Operationen des Geschichtsbewußtseins keine Alternativen darstellen, sondern zwei Seiten ein und derselben Sache« sind und somit vielmehr eine »innere Einheit von beidem, von Erzählen und diskursiver Vernunfttätigkeit, von Imagination und Verstand, von narrativer Sinnbildung und diskursiver Argumentation« besteht. (Rüsen, Jörn: Geschichtsdidaktische Konsequenzen aus einer erzähltheoretischen Historik. In: Siegfried Quandt, Hans Süssmuth (Hg.): Historisches Erzäh-

senschaft ist ein solches Geschichtsverständnis auch Ziel des kompetenzorientierten Geschichtsunterrichts.[57] Studien zeigen jedoch immer wieder, dass dieses Ziel im Geschichtsunterricht nicht erreicht wird.[58] Die Gleichsetzung von Geschichte mit Vergangenheit im Interview ist daher wenig überraschend. Meines Erachtens hat sie jedoch auch Auswirkungen auf den Umgang mit historischen Erzählungen im digitalen Raum und insbesondere auch mit Sprachmodellen, wie ChatGPT.

Auf die Frage, wie der Jugendliche jemandem erklären würde, was ChatGPT eigentlich ist, antwortete er:

> »[…] ist eine KI, also eine künstliche Intelligenz, die da, bei ihr wurde das ganze Internet heruntergeladen, und wenn man eine Frage schickt, sendet, zieht sie die Antwort aus den Informationen, die sie vom Internet hat« (P3_14_m, Pos. 91).

Er besitzt demnach, wenn auch unspezifisch, ein grundlegendes Wissen über die Trainingsdaten der Sprachmodelle, welche sich, wie zuvor erläutert, vor allem aus online verfügbaren Texten zusammensetzen. Dabei verkennt er jedoch, dass keinesfalls alle Daten des Internets hierin eingeschlossen sind und dass diese Daten nur bestimmte Perspektiven und dadurch problematische Verzerrungen beinhalten. Zudem scheint seine Vorstellung von der Funktionsweise der LLM so zu sein, dass diese Informationen aus dem Datensatz »herausziehen« und sie ihm als Antwort auf seine Frage zurücksenden. Dies ist ein problematischer Fehlschluss, denn die Modelle liefern keine Informationen im

---

len. Formen und Funktionen. Göttingen 1982 (Kleine Vandenhoeck-Reihe, Bd. 1485), S. 129–170, hier S. 131).

57 Vgl. Meyer-Hamme, Johannes: Was heißt »historisches Lernen«? Eine Begriffsbestimmung im Spannungsfeld gesellschaftlicher Anforderungen, subjektiver Bedeutungszuschreibungen und Kompetenzen historischen Denkens. In: Thomas Sandkühler u.a. (Hg.): Geschichtsunterricht im 21. Jahrhundert. Eine geschichtsdidaktische Standortbestimmung. Göttingen 2018 (Beihefte zur Zeitschrift für Geschichtsdidaktik, Bd. 17), S. 75–92.

58 Vgl. u.a. die Untersuchungen von Borries, Bodo von: Das Geschichtsbewußtsein Jugendlicher. Erste repräsentative Untersuchung über Vergangenheitsdeutungen, Gegenwartswahrnehmungen und Zukunftserwartungen von Schülerinnen und Schülern in Ost- und Westdeutschland. Weinheim, München 1995; Meyer-Hamme, Johannes: Konzepte von Geschichtslernen und Geschichtsdenken. Empirische Befunde von Schülern und Studierenden (2002). In: Zeitschrift für Geschichtsdidaktik 6 (2007), S. 84–107.

eigentlichen Sinne, sondern willkürliche Wortfolgen auf der Grundlage probabilistischer Berechnungen darüber, wie diese kombiniert werden, aber ohne jeglichen Bezug zur Bedeutung.[59] Daher entstehen auch bei identischem Prompt stets unterschiedliche Ergebnisse und grundsätzliche Fehler, Konfabulationen.[60]

Weiter erläutert der Schüler, dass er und seine Mitschüler*innen das Sprachmodell durchaus auch im Geschichtsunterricht nutzen:

> »[…] wir haben dann ChatGPT nach Informationen zu dem Thema gefragt. Und dann diese genommen und in eigenen Worten einen Satz geschrieben und dann abgegeben« (P3_14_m, Pos. 31–34).

Die Lehrperson hatte zuvor explizit den Auftrag gegeben, ChatGPT hierfür zu nutzen, um »zu schauen, wie es ist« (P3_14_m, Pos. 37). Eine Diskussion über die Ergebnisse fand im Anschluss jedoch nicht statt, stattdessen »gab [es] dann eine Note von der Lehrperson.« (P3_14_m, Pos. 35–36). Leider blieb unklar, wie das Unterrichtssetting genau gestaltet war und zu welchem historischen Thema die Informationen erfragt wurden, da der Junge sich auf Nachfrage nicht genauer erinnern konnte. Grundsätzlich lässt sich hier jedoch zunächst kritisieren, dass ein solcher Unterricht, in dem lediglich historische »Fakten« zusammengetragen und nicht weiter reflektiert wird, welche Deutungen, Perspektiven damit ggf. zusammenhängen, auf welche Quellen sich diese stützen lassen usw., auf ein positivistisches Geschichtsverständnis abzielt. Ebenso wurde die Gelegenheit verpasst, auf die Funktionsweisen und Charakteristiken der Sprachmodelle einzugehen und mit den Lernenden zu reflektieren, inwieweit die Modelle für solche Abfragen überhaupt geeignet sind und was dies für die so entwickelten historischen Erzählungen bedeuten kann, nämlich, dass sie nur bedingt empirisch triftig sind, da sie z.B. nicht auf multiperspektivische Quellen beruhen oder quellenkritische Reflexionen einschließen.[61] Zum einen sind diese Daten nicht in den Trainingsdaten enthalten, zum anderen sind die Modelle nicht in der Lage »historisch Sinn zu bilden« (Jörn Rüsen) oder zu argumentieren. Selbst wenn sie den Anschein

59 Vgl. Bender u.a., On the Dangers of Stochastic Parrots, S. 617.

60 Vgl. Nezhurina u.a., Alice in Wonderland: Simple Tasks Showing Complete Reasoning Breakdown in State-Of-the-Art Large Language Models, S. 8.

61 Zur Frage empirischer Triftigkeit bei Erzählungen in Digitalien vergleiche den Beitrag von Pöppelwiehe in diesem Band.

erwecken, imitieren sie lediglich.[62] Dabei können sie zudem u.a. historische Verschwörungserzählungen oder antisemitische Narrative in den Darstellungen reproduzieren und dadurch zu ihrer Verbreitung beitragen.[63]

Die Jugendlichen nutzen Sprachmodelle zudem auch außerhalb der Schule: » […] dann wollten wir kurz etwas wissen, und dann habe ich kurz nachgeschaut, so auf ChatGPT, weil es einfach eine schnelle und gute Antwort ist.« (P3_14_m, Pos. 97). Auf die Nachfrage, warum es denn eine gute Antwort sei, erklärt er:

> »Es gibt direktere Antworten und mehr so in, ähm, mit Struktur, und du kannst dann immer noch verändern, und so, und auf der Webseite hast du einen fixen Text, und auf ChatGPT kannst du sozusagen, kannst du noch das und das ändern, und dann kannst du das größer machen oder da mehr Informationen oder so« (P3_14_m, Pos. 99).

Da die Modelle in Dialogform mit den Nutzenden agieren, richtet sich das Ergebnis von ChatGPT immer am Prompt aus. Das ist für Nutzende sehr bequem und komfortabel, denn sie müssen nicht erst über eine Suchmaschine nach Inhalten suchen, die mal mehr mal weniger zu ihrer Fragestellung und ihrem Interesse passen. Ebenso vorteilhaft ist es, dass die neusten Versionen der ChatGPT-Modelle zudem die vorhergehenden Prompts und Antworten miteinschließen und daher eine Anpassung, wie oben beschrieben, ermöglichen. Diese Variabilität hängt jedoch vor allem mit der Funktionsweise der Modelle (probabilistische Sequenzen) zusammen. Auch wenn der Jugendliche dies nicht erkennt, ist sein Umgang mit dem Sprachmodell nicht völlig unkritisch, denn er reflektiert, dass er nur schwer überprüfen kann, ob der Inhalt anhand von Quellen belegbar ist und woher er stammt: »Das kann man nicht genau beurteilen, weil man weiß ja nicht, wo die Quelle her ist« (P3_14_m, Pos. 103). Und auch im Vergleich zu historischen Darstellungen auf Webseite fallen ihm Unterschiede auf:

---

62 Vgl. hierzu die zuvor beschriebenen Untersuchung von Nezhurina u.a. (Nezhurina u.a.: Alice in Wonderland: Simple Tasks Showing Complete Reasoning Breakdown in State-Of-the-Art Large Language Models.

63 Vgl. Lin, Stephanie; Hilton, Jacob; Evans, Owain, 2021: TruthfulQA: Measuring How Models Mimic Human Falsehoods, http://arxiv.org/pdf/2109.07958v2, aufgerufen am 14.02.2024.

> »Inhaltlich ist es schwierig zu sagen. Ich denke, es kommt ja beides, ursprünglich von der Webseite. ChatGPT hat ja auch die Informationen von der Webseite, und ich denke, die Informationen sind ähnlich. Aber auf der Webseite findest du zwar mehr zu einem bestimmten Thema, dafür wahrscheinlich mehr richtig und auf ChatGPT, ähm, verschiedene Antworten.« (P3_14_m, Pos. 101)

Er erkennt also, dass die Ergebnisse von ChatGPT keinen überprüfbaren Beleg liefern und dass dies oftmals einen Unterschied zu historischen Darstellungen, etwa auf Wikipedia darstellt, wo Quellen- und Literaturnachweise aufgeführt werden. Nichtsdestotrotz bestimmt diese Erkenntnis nicht seinen alltäglichen Umgang mit dem Large Language Model: »Dann weiß man natürlich nicht immer, ob es stimmt, aber ich vertraue dem« (P3_14_m, Pos. 87). Was er später auch nochmals bestärkt: »Ja, meistens vertraue ich dem und denke, dass es richtig ist« (P3_14_m, Pos. 105).

5 Fazit und Ausblick

Grundsätzlich lässt sich zeigen, dass KI-generierte historische Erzählungen die Gefahr bergen, Nutzer*innen zu überwältigen, da sie »Objektivität« und Sprachverstehen vortäuschen. Tatsächlich beziehen die Trainingsdaten, wie im ersten Teil des Beitrags ausführlich beschrieben, aber nur ganz bestimmte Inhalte, Perspektiven und Deutungen (sowie auch Verzerrungen und Diskriminierungen) mit ein, welche damit vor allem auch die normative Ebene der Erzählungen beeinflussen. Gerade Perspektiven marginalisierter Gruppen sind oftmals nicht in den Trainingsdaten enthalten und damit auch nicht in den historischen Erzählungen.[64] Jugendliche können den Eindruck gewinnen, sie würden durch die LLM eine historische »Supererzählung« erhalten, da das Modell vermeintlich auf das gesamte »Wissen« (vgl. Interview: »das ganze Internet«) zugreift und dieses sogar adressatengerecht und überzeugend aufbereitet. Ebenso problematisch ist die unzureichende empirische Triftigkeit, welche dadurch entsteht, dass es nicht nachvollziehbar ist, woher die einzelnen Elemente und Wortfolgen der Erzählung genau stammen. Außerdem können die Narrationen durch die zufällige Zusammensetzung von Wortfolgen

64 Vgl. hierzu auch den Beitrag von Anja Neubert in diesem Band.

Konfabulationen und Fehler produzieren, die so noch nicht einmal in den Trainingsdaten enthalten sind. Durch das grundsätzlich fehlende Sprachverständnis kann auch nicht davon ausgegangen werden, dass die KI-Darstellungen narrativ triftig sind, d.h. dass sie in sich stimmig sind.

Was bedeutet dies nun aber für historische Lernprozesse im Unterricht? M.E. sollten genau diese Aspekte und Überlegungen Teil eines kompetenzorientierten Geschichtsunterrichts sein. KI-generierte Erzählungen prägen z.B. auf Social Media oder auch grundsätzlich im digitalen Raum die Alltagswelt der Lernenden. Sie sollten also Kompetenzen erwerben, besonders im Bereich der De-Konstruktion, um mit diesen kritisch-reflektiert umgehen zu können und sich zu ihnen zu verhalten.

Erste pragmatische Unterrichtsideen und Konzepte wurden bereits publiziert. Klar und entschieden abzulehnen sind davon jedoch jene Beispiele, welche eine Art historisches Rollenspiel mit den Sprachmodellen vorschlagen: Die Lernenden sollen z.B. einen Dialog mit Otto von Bismarck mit dem Modell nachspielen, eine »digitale Zeitreise« unternehmen, in welcher das LLM die Rolle der jeweiligen historischen Persönlichkeit übernimmt. Ein solches Setting fördert ein falsches Verständnis über die KI-Modelle als auch ein positivistisches Geschichtsverständnis (d.h.: als könnten wir in der Zeit zurückreisen und Geschichte ist das, was damals passiert ist).[65]

Weitaus überzeugendere Beispiele finden sich dagegen z.B. im Praxisheft von Oliver Held: Hier werden in zahlreichen Vorschlägen historische Erzählungen von Historiker*innen mit KI-generierten gegenübergestellt und einer eingehenden Analyse unterzogen, ähnliche Beispiele finden sich für historische und KI-erstellte Quellen sowie Analysen. Lediglich zu kurz kommt in den Beispielen und Arbeitsmaterialien eine genauere Erklärung, wie die Sprachmodelle funktionieren und welche Auswirkungen dies auf die Ergebnisse haben kann.[66] Dies ließe sich jedoch leicht anhand der bereits vielfältigen Handreichungen aus dem Bereich Medienbildung zu KI und LLM ausbessern. Empfehlenswert sind z.B. die Materialien Annabel Lindner und Stefan Seegerer »AI

65 Vgl. Mayer, Thomas: Chatten mit historischen Persönlichkeiten. Kompetenzcheck mit ChatGPT. In: Geschichte lernen (2023), H. 213, S. 62f.

66 Vgl. Held, Oliver: ChatGPT im Geschichtsunterricht. Frankfurt 2024 (Geschichte unterrichten).

Unplugged. Wir ziehen künstlicher Intelligenz den Stecker«, welche spielerisch an ein kritisch-reflektiertes Verständnis von KI heranführen.[67]

Somit bleibt es heute noch viel mehr unsere gesellschaftliche Aufgabe, uns stets der »Fehlbarkeit und Begrenztheit der Systeme«[68] bewusst zu sein und diese auch beim Umgang mit KI-generierten historischen Narrationen zu bedenken.

Literatur

Barera, Michael, 2020: Mind the Gap: Addressing Structural Equity and Inclusion on Wikipedia, http://hdl.handle.net/10106/29572, aufgerufen am 14.02.2024.

Bender, Emily M; Gebru, Timnit; McMillan-Major, Angelina; Shmitchell, Shmargaret: On the Dangers of Stochastic Parrots. In: Conference on Fairness, Accountability, and Transparency (FAccT ›21), March 3–10, 2021, Virtual Event, Canada (2021), S. 610–623.

Borries, Bodo von: Das Geschichtsbewußtsein Jugendlicher. Erste repräsentative Untersuchung über Vergangenheitsdeutungen, Gegenwartswahrnehmungen und Zukunftserwartungen von Schülerinnen und Schülern in Ost- und Westdeutschland. Weinheim und München 1995.

Held, Oliver: ChatGPT im Geschichtsunterricht. Frankfurt 2024 (Geschichte unterrichten).

Hielscher, Michael: SoekiaGPT – ein didaktisches Sprachmodell. In: Informatische Bildung in Schulen 1 (2023), H. 1, S. 1–11.

Hodel, Jan: Verkürzen und Verknüpfen. Geschichte als Netz narrativer Fragmente: Wie Jugendliche digitale Netzmedien für die Erstellung von Referaten im Geschichtsunterricht verwenden. Bern 2013 (Geschichtsdidaktik heute, Bd. 5).

Hubinger, Evan u.a., 2024: Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training, http://arxiv.org/pdf/2401.05566v3, aufgerufen am 14.02.2024.

Krebs, Alexandra: Geschichten im digitalen Raum. Historisches Lernen in der »App in die Geschichte«. Berlin 2024 (Medien der Geschichte, Bd. 7).

67 Vgl. Lindner, Annabel; Seegerer, Stefan: AI Unplugged. Wir ziehen künstlicher Intelligenz den Stecker, www.aiunplugged.org/german.pdf, aufgerufen am 14.02.2024.

68 Weizenbaum, Alptraum Computer.

Lin, Stephanie; Hilton, Jacob; Evans, Owain, 2021: TruthfulQA: Measuring How Models Mimic Human Falsehoods, http://arxiv.org/pdf/2109.07958v2, aufgerufen am 14.02.2024.

Lindern, Jakob von; Wolfangel, Eva, 2024: Ist das der Papst?, https://www.zeit.de/2024/13/diversitaet-google-ki-gemini-bild-generator-papst, aufgerufen am 14.02.2024.

Lindner, Annabel; Seegerer, Stefan: AI Unplugged. Wir ziehen künstlicher Intelligenz den Stecker, www.aiunplugged.org/german.pdf, aufgerufen am 14.02.2024.

Maslej, Nestor u.a.: The AI Index 2024 Annual Report. AI Index Steering Committee, Institute for Human-Centered AI. Stanford 2024.

Mayer, Thomas: Chatten mit historischen Persönlichkeiten. Kompetenzcheck mit ChatGPT. In: Geschichte lernen (2023), H. 213, S. 62f.

Meyer-Hamme, Johannes: Konzepte von Geschichtslernen und Geschichtsdenken. Empirische Befunde von Schülern und Studierenden (2002). In: Zeitschrift für Geschichtsdidaktik 6 (2007), S. 84–107.

Meyer-Hamme, Johannes: Was heißt »historisches Lernen«? Eine Begriffsbestimmung im Spannungsfeld gesellschaftlicher Anforderungen, subjektiver Bedeutungszuschreibungen und Kompetenzen historischen Denkens. In: Sandkühler, Thomas u.a. (Hg.): Geschichtsunterricht im 21. Jahrhundert. Eine geschichtsdidaktische Standortbestimmung. Göttingen 2018 (Beihefte zur Zeitschrift für Geschichtsdidaktik, Bd. 17), S. 75–92.

Nezhurina, Marianna; Cipolina-Kun, Lucia; Cherti, Mehdi; Jitsev, Jenia: Alice in Wonderland, 2024: Simple Tasks Showing Complete Reasoning Breakdown in State-Of-the-Art Large Language Models, https://arxiv.org/pdf/2406.02061, aufgerufen am 16.07.2024.

Nicoletti, Leonardo; Bass, Dina, 2023: Humans Are Biased. Generative AI is even worse, https://www.bloomberg.com/graphics/2023-generative-ai-bias/, aufgerufen am 14.02.2024.

O.A.: Common Crawl, https://commoncrawl.org/, aufgerufen am 14.02.2024.

O.A.: Epoch AI, https://epochai.org/, aufgerufen am 14.02.2024.

O.A.: Stable Diffusion, https://stablediffusionweb.com/de, aufgerufen am 14.02.2024.

O.A.: WebText, https://paperswithcode.com/dataset/webtext, aufgerufen am 14.02.2024.

OpenAI u.a., 2023: GPT-4 Technical Report, https://arxiv.org/pdf/2303.08774, aufgerufen am 14.02.2024.

Ouyang, Long u.a., 2022: Training language models to follow instructions with human feedback, https://arxiv.org/pdf/2203.02155, aufgerufen am 14.02.2024.

Rosenfeld, Ronald: Two decades of statistical language modeling: Where do we go from here? In: Proceedings of the IEEE 88 (2000), S. 1270–1278.

Rüsen, Jörn: Geschichtsdidaktische Konsequenzen aus einer erzähltheoretischen Historik. In: Quandt, Siegfried; Süssmuth, Hans (Hg.): Historisches Erzählen. Formen und Funktionen. Göttingen 1982 (Kleine Vandenhoeck-Reihe, Bd. 1485), S. 129–170.

Schreiber, Waltraud: Kompetenzbereich historische Methodenkompetenz. In: Körber, Andreas; Schreiber, Waltraud; Schöner, Alexander (Hg.): Kompetenzen historischen Denkens. Ein Strukturmodell als Beitrag zur Kompetenzorientierung in der Geschichtsdidaktik. Neuried 2007 (Kompetenzen: Grundlagen – Entwicklungen – Förderung, Bd. 2), S. 194–235.

Schwarz, Karolin: Hasskrieger. Der neue globale Rechtsextremismus. Sonderausgabe für die Bundeszentrale für politische Bildung. Bonn 2020 (Schriftenreihe/Bundeszentrale für Politische Bildung, Band 10545).

Shanahan, Murray: Talking about Large Language Models. In: Commun. ACM 67 (2024), H. 2, S. 68–79.

Walter, Yoshija: Die Vermenschlichung der künstlichen Intelligenz. Die grosse sozio-psychologische Kritik und das KI-Bewusstsein 2023, https://www.kalaidos-fh.ch/de-CH/Blog/Posts/2023/11/Digitalisierung-1121-Vermenschlichung-kuenstliche-Intelligenz, aufgerufen am 14.02.2024.

Weizenbaum, Joseph, 1972: Alptraum Computer, https://www.zeit.de/1972/03/alptraum-computer, aufgerufen am 14.02.2024.

Weizenbaum, Joseph: Computer Power and Human Reason. From Judgement to Calculation. San Francisco 1976.

Westra, Laura; Lawson, Bill E.: Faces of environmental racism. Confronting issues of global justice. 2. Aufl. Lanham 2001 (Studies in social, political, and legal philosophy).

Wineburg, Sam: Warum historische Kompetenzen für die Auswertung von digitalen Quellen nicht ausreichend sind. In: Barsch, Sebastian; Lutter, Andreas; Meyer-Heidemann, Christian (Hg.): Fake und Filter. Historisches und politisches Lernen in Zeiten der Digitalität. Frankfurt a.M. 2019 (Wochenschau Wissenschaft), S. 105–120.