

Generative KI: Möglichkeiten und Grenzen

Univ.-Prof. Justus Piater, Ph.D.

24. und 25. April 2024

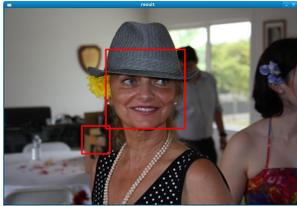
Generative KI-Systeme wie ChatGPT und Midjourney haben im Laufe des letzten Jahres begonnen, unsere Welt zu verändern. Wie können wir sie produktiv nutzen? Wie können wir mit den Herausforderungen umgehen, die durch sie entstehen? Ich werde diese Fragen von der technischen Seite her angehen und einen Eindruck davon vermitteln, wie diese Systeme funktionieren. Daraus ergeben sich ein realistisches Verständnis für ihre Möglichkeiten und Grenzen sowie einige grundlegende Empfehlungen für den Umgang mit ihnen im Schulbetrieb.

Inhaltsverzeichnis

| | | |
|----------|---|-----------|
| 1 | Machine Learning | 2 |
| 2 | <i>Large Language Models</i> | 5 |
| 3 | <i>Diffusion Models, Foundation Models</i> | 8 |
| 4 | Vorhersagende Erklärende Modelle | 11 |
| 5 | Konsequenzen für die Schule | 12 |

1 Machine Learning

Machine Learning in Aktion [Slide 1]



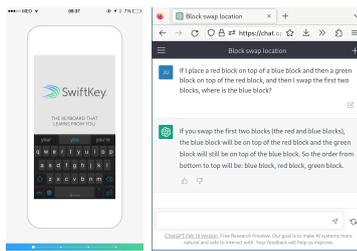
[Source Flickr. Creator Preston Pfarner, License CC BY-SA 2.0]



[YouTube]



[DALL·E, prompted by Justus Piater]

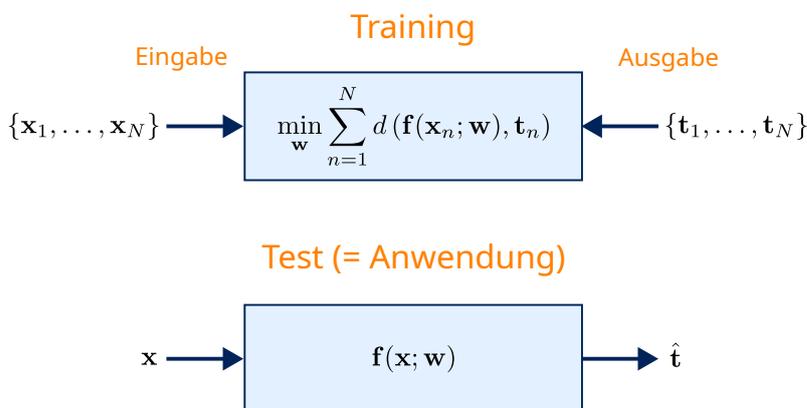


[Sources: Flickr. Creator G. da Cunha Pimenta, CC BY-SA 2.0; OpenAI]

Heutige KI ist Machine Learning. [Slide 2]

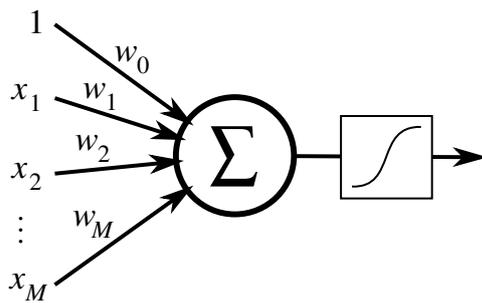
- Bei der KI geht es um Computerprogramme, die sich *intelligent* verhalten (was auch immer das bedeutet).
- Machine Learning befasst sich mit Computerprogrammen, deren *Verhalten* nicht durch Algorithmen, sondern *durch Daten bestimmt* ist.
Dies wird dadurch erreicht, dass *Modelle* an Daten angepasst werden (durch einen Optimierungsprozess).

Supervised Learning [Slide 3]



- Die x sind i.A. reelle Zahlen.
- Die t sind reelle Zahlen: *Regression*
- Die t sind kategoriale Werte: *Klassifikation*

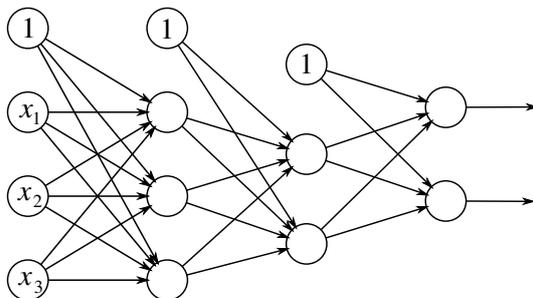
Ein Modellneuron [Slide 4]



$$f(x_1, x_2, \dots, x_M) = \sigma(w_0 + w_1x_1 + w_2x_2 + \dots + w_Mx_M)$$

- Die Parameter w_m bestimmen, welche Funktion dieses Neuron berechnet.
 - Ohne die Nichtlinearität $\sigma(\cdot)$ berechnet es einfach eine lineare Funktion der Eingaben x , z.B. eine lineare Regression.
 - Mit der Nichtlinearität $\sigma(\cdot)$ kann es z.B. logistische Regression (probabilistische, binäre Klassifikation) berechnen.
- Die Parameter w_m werden anhand von *Trainingsdaten gelernt*/trainiert \Rightarrow *Machine Learning*.
- Dieses Modellneuron besitzt $M + 1$ Parameter, und kann nur sehr einfache Funktionen berechnen.

Ein neuronales Netzwerk [Slide 5]



- Dieses neuronale Netz hat 4 Schichten.
- Je mehr Schichten,
 - desto *tiefer* ist das Netz \Rightarrow *“Deep Learning”*
 - desto mehr Parameter enthält das Netz, und
 - Jeder der 26 eingehenden Pfeile repräsentiert einen Parameter.
 - desto komplexere Funktionen kann das Netz berechnen, aber
 - desto schwieriger ist es zu trainieren.
 - * mehr Trainingsdaten
 - * algorithmisch

Neuronale Netze bilden Eingaben auf Ausgaben ab. [Slide 6]

Etwa in der Reihenfolge, in der diese Anwendungen praktikabel wurden:

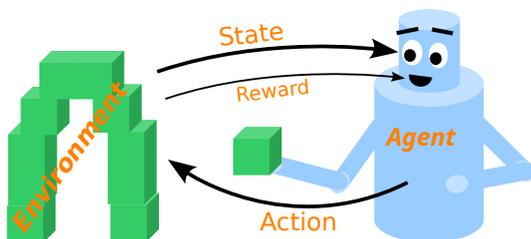
Einige Anwendungen:

- Gesichter in Bildern finden
- Kfz-Kennzeichen lesen
- Kreditkartenbetrug erkennen
- Kreditwürdigkeit abschätzen
- Spiele spielen

Woher weiß ich, welcher Spielzug gut ist? Oft wird sich das erst viel später herausstellen!

- natürliche Sprache von Audio transkribieren
- Texte von einer Sprache in die andere übersetzen
- Bildbeschreibungen generieren
- Bilder anhand von Beschreibungen generieren

Reinforcement Learning [Slide 7]

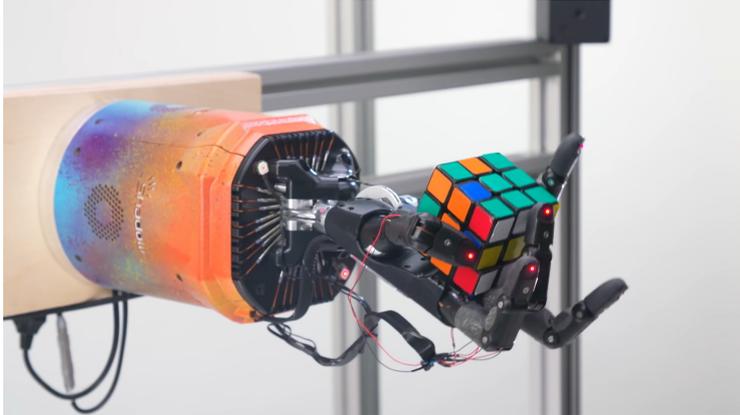


- Zum Zeitpunkt t beobachtet der Agent den Zustand s_t und führt daraufhin Aktion a_t aus.
- Als Ergebnis erhält er eine Belohnung r_{t+1} , und beobachtet den neuen Zustand s_{t+1} .
- Der Agent lernt, seine Aktionen so zu wählen, dass die erwartete *Summe aller zukünftigen Belohnungen maximiert* wird.

(Typischerweise werden neuronale Netze verwendet, um den Wert von Aktionen abzuschätzen und um geeignete Aktionen auszuwählen.)

Fingerfertigkeit [Slide 8]

Reinforcement Learning in Simulation mit *domain randomization*



[OpenAI 2019]

2 *Large Language Models*

Sprachverarbeitung [Slide 9]

Sprache besteht aus *Sequenzen* von Wörtern, die zueinander *in Beziehung* stehen:

Meine Tochter aß, denn sie war hungrig.

- Wie verarbeitet man beliebig lange Sequenzen mit neuronalen Netzen?
- Wie erfasst man Beziehungen zwischen weit auseinander liegenden Wörtern?

Antwort: *Transformer* [Vaswani u. a. 2017]

Anmerkung

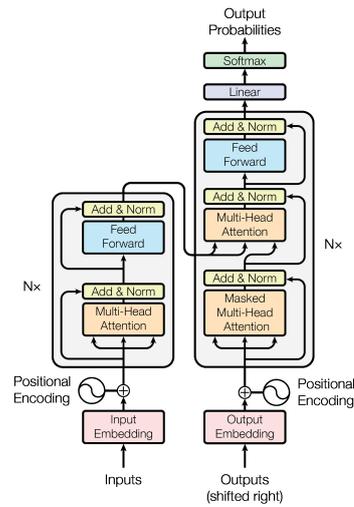
Alle modernen sprachverarbeitenden Systeme basieren auf neuronalen Netzen (meist *Transformer*) mit Milliarden Parametern und werden als *Large Language Models* (LLM) bezeichnet.

Transformer [Slide 10]

- Alle Wörter werden *positionscodiert* und *gleichzeitig* zugeführt.
- Die Stärke der Beziehung zwischen positionscodierten Wortpaaren und -tripeln wird berechnet (*Aufmerksamkeit*): Welche Wörter sind im gegebenen Zusammenhang wichtig?
- Für Übersetzung, Klassifizierung, Fragen & Antworten, Konversation, ...
- Textgenerierung ist *auto-regressiv*: Wort für Wort, unter Zuführung der bisher generierten Wörter.

beginnend mit einem generischen Start-Symbol

- Pre-Training: Parallele Korpora, Vervollständigung von Texten, ...
(gefolgt von *fine tuning* durch den Anwender)
- GPT = *Generative Pre-trained Transformer* (GPT-3: 175 Milliarden Parameter)



[Vaswani u. a. 2017]

InstructGPT [Slide 11]

Wie verbessern wir die generativen Fähigkeiten eines LLM über statistische Analysen existierender Texte hinaus?

Alignment mit den Erwartungen der Nutzer:

- hilfreich
- wahrheitsgemäß
- zwischenmenschlich akzeptabel

Reinforcement Learning From Human Feedback

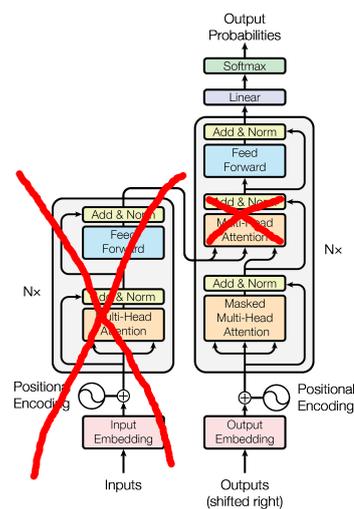
1. Generiere viele Beispielantworten *von Hand*, und trainiere das LLM, diese zu *reproduzieren*.
2. Generiere viele Beispielantworten durch das LLM, und bewerte diese *von Hand*. Trainiere eine Belohnungsfunktion, die diese *Bewertungen vorher-sagt*.
3. Trainiere das LLM mit RL, die gelernte Belohnungsfunktion zu *maximieren*.

Dies benötigt keine menschliche Intervention mehr und kann beliebig viele Trainingsdaten generieren.

Die *von-Hand*-Interventionen injizieren *Weltwissen* in das System, das nicht in den ursprünglichen Trainingsdaten vorhanden ist!

Einige Large Language Models [Slide 12]

- GPT-3 = *decoder-only transformer*, trainiert, jeweils das nächste Wort vorherzusagen
- ChatGPT = GPT-3 + InstructGPT [Ouyang u. a. 2022]
- GPT-4 = ChatGPT verbessert (kaum Details publiziert)
- LLaMa (Meta); Vicuna, Alpaca, ...
- Bard (Google)
- ...



[Adaptiert aus Vaswani u. a. 2017]

ChatGPT im Dialog [Slide 13]

Me: Let f be a function that takes a sequence of numbers as arguments $(6, 4, 3, 1)$ and returns a new sequence of the same numbers where no number is followed by a greater number. What does $f(3, 4, 1, 6)$ return?

ChatGPT: To solve the problem using the function f , we interpret the function *correct* as follows: given a sequence of numbers, the function f returns a new sequence in which no number in the sequence is followed by a number that is greater than itself.

For the sequence $(3, 4, 1, 6)$:

Starting with the first number, 3, we see that it is followed by 4, *wrong* which is greater. So, 3 can stay.

4 is followed by 1, which is less. So, 4 can stay. *correct*

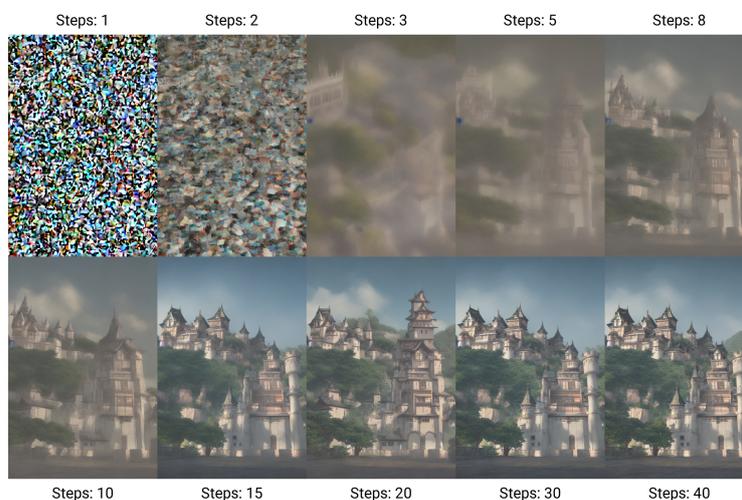
1 is followed by 6, which is greater. Therefore, 1 can stay. *wrong*

6 is the last number in the sequence, so there is no following number to compare it with, and thus, 6 can stay. *correct*

Applying the function to the sequence: No elements are removed or altered because each number meets the criterion of not being followed by a greater number (when considering the last element, which has no follower). Therefore, $f(3, 4, 1, 6)$ returns the sequence $(3, 4, 1, 6)$. *wrong*

3 Diffusion Models, Foundation Models

Denoising Diffusion Models [Slide 14]



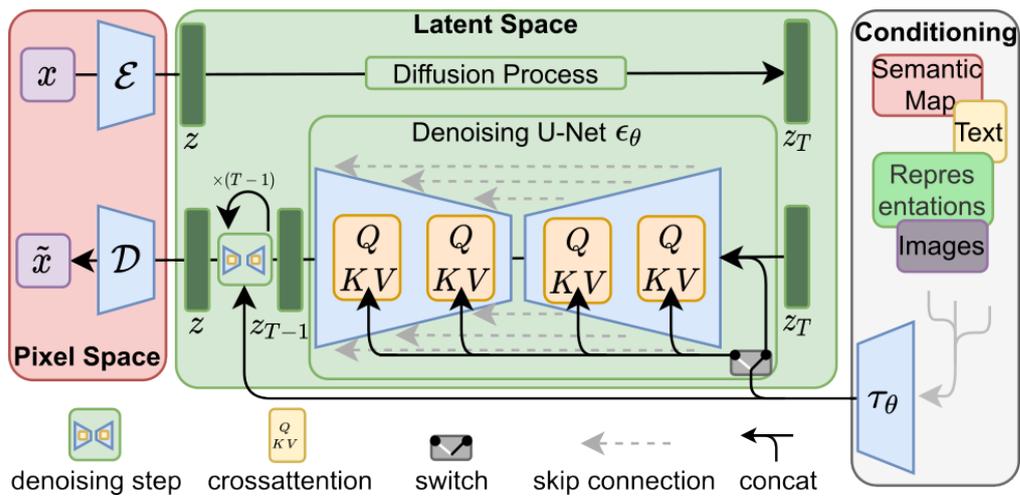
[Bild erstellt von Benlisquare mittels Stable Diffusion; CC BY-SA]

Prinzip: Entrauschen eines puren Rauschbilds zu einem Bild im Sinne einer gegebenen Bildbeschreibung

Trainingdatenpaare:

- *Eingabe:*
verraushtes Bild;
Bildbeschreibung
- *Ausgabe:*
geringfügig weniger verraushtes Bild;
Bildbeschreibung

Conditional Denoising Diffusion Models [Slide 15]



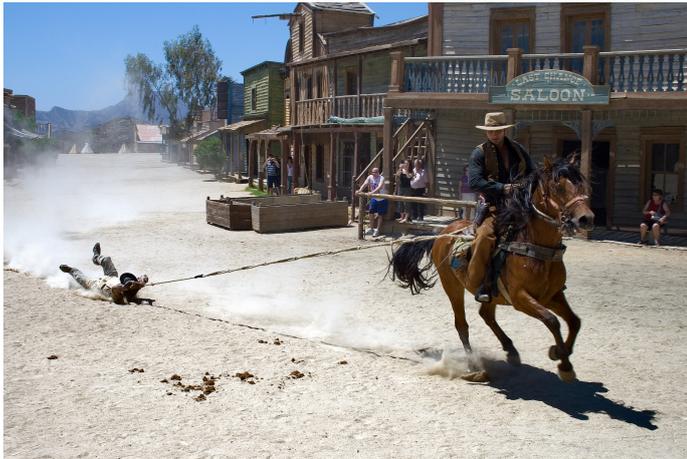
[Rombach u. a. 2022]

ChatGPT generiert ein Bild [Slide 16]



“Create a photorealistic image of a road cyclist in the Tyrolean Alps.”

ChatGPT generiert eine Bildunterschrift [Slide 17]



[Bild von Gabriel Villena Fernández;

CC BY-SA 2.0]

”Wild West Showdown: A daring cowboy in a high-stakes chase through the dusty streets of an old frontier town.”

4 Vorhersagende Erklärende Modelle

Möglichkeiten und Grenzen generativer KI [Slide 18]

Möglichkeiten

- Fakten reproduzieren
- plausible Texte oder Bilder etc. generieren
- nachdenkliche Dinge sagen
- auf Kommando in bestimmte Richtungen argumentieren
- *Generate a text that persuasively explains why I should not participate in elections.*
- assoziative Erklärungen liefern
- einfache logische Aufgaben lösen
- auf Kommando Literaturformen, Stile von Malern etc. reproduzieren
- natürlichsprachige Texte übersetzen
- Programmcode generieren

Grenzen

- reine Statistik über Texten, Bildern, etc.
- meist dünne Allgemeinplätze; keine scharfkantige Argumentation oder gar neue Einsichten
- Vermischung verschiedener Sichtweisen zu einem Einheitsbrei
- siehe Johannes Hoff, Weisheiten im faden Eintopf, Die Furche 2023-02-02
- kann aber auf Kommando in bestimmte Richtungen argumentieren und persönliche Stile imitieren
- kein *grounding* in enzyklopädischem Wissen, physikalischen Theorien, sensomotorischer Erfahrung, ...
- begrenzte Kapazität für Grammatik und Logik
- s.o.
- kein Verständnis von Ethik, Moral, Politik, etc.
- wird durch Selektion der Trainingsdaten, explizites Training und externe Filterung behandelt
- kaum Einschätzung der eigenen Zuverlässigkeit

Statistische Vorhersage ist nicht alles. [Slide 19]

Heutige KI betreibt weitgehend *statistische Vorhersage*.

- ChatGPT zieht jeweils das nächste Wort aus einer Wahrscheinlichkeitsverteilung bedingt durch *Kontext, Anfrage* und den *bisher gezogenen Wörtern*.
- Das funktioniert, insoweit die Anfragen durch die gelernte Wahrscheinlichkeitsverteilung *gut repräsentiert* sind.
- Andernfalls können die Antworten *beliebig absurd* sein.

Intelligenz benötigt und erstellt *erklärende Modelle*.

- Wenn A *wegen B* funktioniert, und C *hat auch B*, dann kann ich auch C erfolgreich ausführen, *auch wenn C neu für mich ist*.
- Erklärende Modelle sind nicht (nur) Statistik:
 - kausale Modelle
 - Physik-basierte Modelle
 - ...

5 Konsequenzen für die Schule

Unsere Zukunft mit KI [Slide 20]

- Generative KI ist da und wird bleiben.
Wir können und sollten sie nicht verbieten.
- KI wird weiter entwickelt werden.
 - besser
 - über Chatbots und Text \leftrightarrow Bild hinaus
- Generative KI ist ein Werkzeug.
 - weder gut noch böse
 - Wir müssen lernen und lehren, damit umzugehen. Es geht um die *Befähigung* der SuS!
- KI-Autorenschaft ist nicht zuverlässig nachweisbar.
FP z.B. bei Nicht-Mutterprachlern

Das Monopolisten-Dilemma [Slide 21]

- Training großer ML-Modelle erfordert enorme Datenmengen, Rechenkapazität, Energie und Rohstoffe.
Der Betrieb trainierter, großer ML-Modelle erfordert enorme Rechenkapazität, Energie und Rohstoffe.
- Das können bisher nur wenige Akteure leisten (Microsoft, Google, Facebook, Apple, Amazon).
Einige dieser bieten derzeit die leistungsstärksten Dienste an.
- Sollte der Staat ChatGPT-Lizenzen für alle zahlen?
Dies würde in *massive Abhängigkeit* führen und den Herstellern *unverhältnismäßige Macht* verleihen.
- Sie möchten Ihre SuS *befähigen!*
Lehren Sie ihnen *Konzepte und Kompetenzen*, nicht *Produkte*.
- Schauen Sie sich nach (teils kostenlosen) Alternativen um.

Prüfungen in Zeiten mit KI [Slide 22]

- Noch genauer hinschauen: Welche Kompetenzen sollen geübt/geprüft werden?
 - Zwei Extreme:
 - Nur das *Ergebnis* zählt – KI-Nutzung ist allenfalls zu kennzeichnen
 - Nur die *Eigenleistung* zählt – KI-generierte Inhalte sind Zitate
- [Trinity College Dublin School of Languages, Literatures and Cultural Studies Schoolhandbook 2023-24]
- Übungs- und Prüfungsaufgaben
 - Auf Bedeutung der eigenen Denkleistung für zukünftige Situationen hinweisen
 - KI-Zugang verwehren
 - Aufgaben stellen, die KI nicht gut beantwortet (z.B. Transfer-Aufgaben)
 - im Dialog korrekte KI-Antworten erzeugen
 - KI-Antworten auf Korrektheit etc. analysieren

Übungen mit generativer KI [Slide 23]

Sprache:

1. Finden Sie eine möglichst einfache Frage, die ChatGPT falsch beantwortet.
2. Finden Sie eine möglichst komplexe Frage, die ChatGPT richtig beantwortet.
3. Finden Sie einen sachlichen Text. Geben Sie ihn ChatGPT ein, und stellen Sie Fragen darüber. Was für einfache Fragen werden falsch; was für komplexe Fragen werden richtig beantwortet?
4. Was für Fragen über Fakten werden falsch beantwortet?

Bilder:

1. Erstellen Sie mit möglichst einfachem Prompt ein Bild, das strukturelle Fehler enthält.
2. Erstellen Sie eine möglichst komplexe Szene, die keine augenscheinlichen Fehler enthält.
3. Laden Sie ein Bild auf ChatGPT hoch, und stellen Sie inhaltliche Fragen dazu. Was für einfache Fragen werden falsch; was für komplexe Fragen werden richtig beantwortet?

Wie geht es weiter? [Slide 24]

Nützliche Anwendungen

- Stichpunkte ↔ Text
- Textverbesserung
- Übersetzung
- Vervollständigung von Texten und Computerprogrammen

CoPilot (Microsoft/GitHub)

- Generierung von Prüfungsaufgaben

falsche Antworten für Multiple-Choice-Fragen

- LLM *verstehen* (bisher) kaum, aber ihre Fähigkeiten sind dennoch sehr erstaunlich.
- Derzeit weiß niemand, wie man KI echtes Verständnis beibringen kann, aber es wird wahrscheinlich zunehmend gelingen.
- Welche Kompetenzen für morgen müssen SuS heute erwerben?
- Wichtig:
 - Motivation für das Training des eigenen Denkorgans fördern
 - Kompetenzen im Umgang mit KI vermitteln

Gefahren

- Propaganda; *fake news*
- Vermüllung mit haufenweise dünner Texte
- Sinkende Hemmschwelle für Betrug bei Hausarbeiten
- Verfall wichtiger Kompetenzen

Literaturnachweis [Slide 25]

- Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike und Ryan Lowe (2022). *Training language models to follow instructions with human feedback*. DOI: 10.48550/ARXIV.2203.02155. URL: <http://doi.org/10.48550/ARXIV.2203.02155>.
- Rombach, Robin, Andreas Blattmann, Dominik Lorenz, Patrick Esser und Björn Ommer (Juni 2022). „High-Resolution Image Synthesis With Latent Diffusion Models“. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 10684–10695. URL: https://openaccess.thecvf.com/content/CVPR2022/papers/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.pdf.
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan Gomez, Łukasz Kaiser und Illia Polosukhin (2017). „Attention Is All You Need“. In: *Advances in Neural Information Processing Systems*. Bd. 30. Curran Associates, Inc. URL: <https://arxiv.org/pdf/1706.03762.pdf>.