

Was ist und wie funktioniert ChatGPT?

16. Jenaer Medienrechtliche Gespräche

Prof. Dr. Clemens Beckstein

Friedrich-Schiller-Universität Jena

9. November 2023

Was ist ChatGPT

Ein Software-Werkzeug,
das „**Texte**“ assoziiert.

Es generiert

- frei formulierte,
„**passende Antworten**“ auf
- frei formulierte
Eingabetexte (**Prompts**).

ChatGPT ist **kein Google**
(und hilft hier seltsam weiter fort)!



Finde eine Fortsetzung für diesen Text:

Geschrieben steht: »Im Anfang war das Wort!«
Hier stock ich schon! Wer hilft mir weiter fort?



Geschrieben steht: »Im Anfang war das Wort!«
Hier stock ich schon! Wer hilft mir weiter fort?

Ich kann nicht das hohe Geheimnis weisen,
Von dem die alten Meister leise preisen.

Doch wag ich mich mit sinnigem Bedacht,
Zu suchen nach der Wahrheit tiefer Macht.

Im Anfang war der Sinn, der alles trägt,
Das Wort erst Leben durch den Sinn erhält.

Johannes 1:1, Neue-Welt-Übersetzung:

»Am Anfang war das Wort, und das Wort war bei Gott, und das Wort war ein Gott.«

Goethe im Studierzimmer-Monolog des **Faust 1**:

*»Geschrieben steht: „Im Anfang war das Wort!“ Hier stock ich schon!
Wer hilft mir weiter fort? Ich kann das Wort so hoch unmöglich schätzen,
Ich muß es anders übersetzen, Wenn ich vom Geiste recht erleuchtet bin.*

*Geschrieben steht: Im Anfang war der Sinn. Bedenke wohl die erste Zeile,
Daß deine Feder sich nicht übereile! Ist es der Sinn, der alles wirkt und schafft?*

*Es sollte stehn: Im Anfang war die Kraft! Doch, auch indem ich dieses niederschreibe,
Schon warnt mich was, daß ich dabei nicht bleibe. Mir hilft der Geist!
Auf einmal seh ich Rat Und schreibe getrost: Im Anfang war die Tat!«*



Jorge Luis Borges „Bibliothek von Babel“

Wie findet man Wissen
in der Flut ($10^{1.845.281}$) aller **erdenklicher** Bücher?

Gibt es „**Schlüsselbücher**“,
die viele andere Bücher in einen Kontext setzen?
Ohne die **Bibliothekare** ist die Bibliothek nichts!

(Borges 1992)

Algorithmische Sinn-Suche

Digitalisierung der Bibliothek als riesige Assoziations-Tabelle:

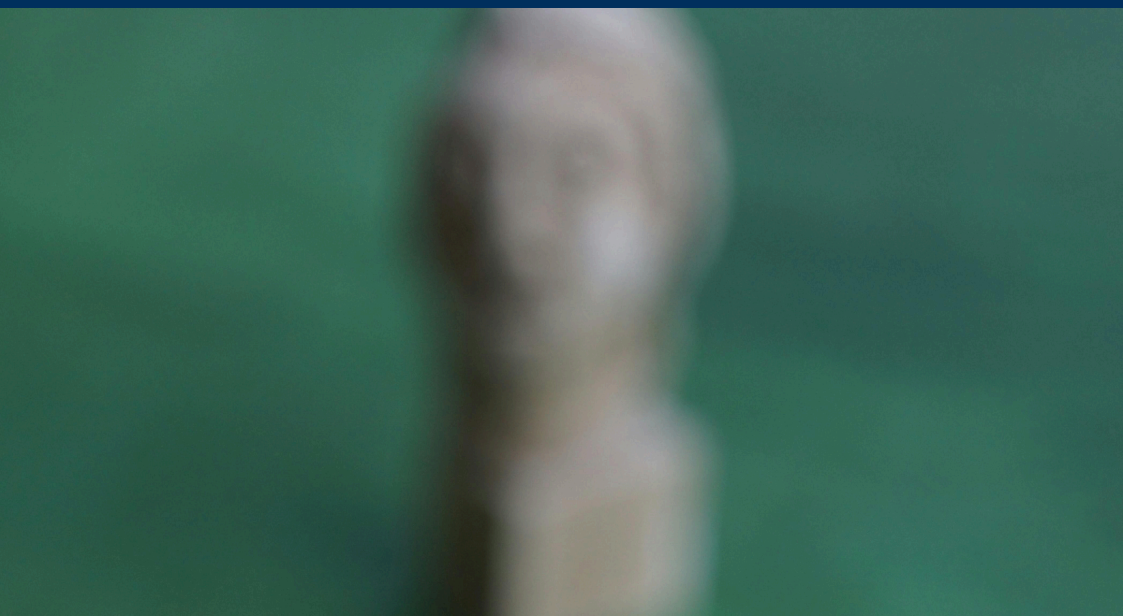
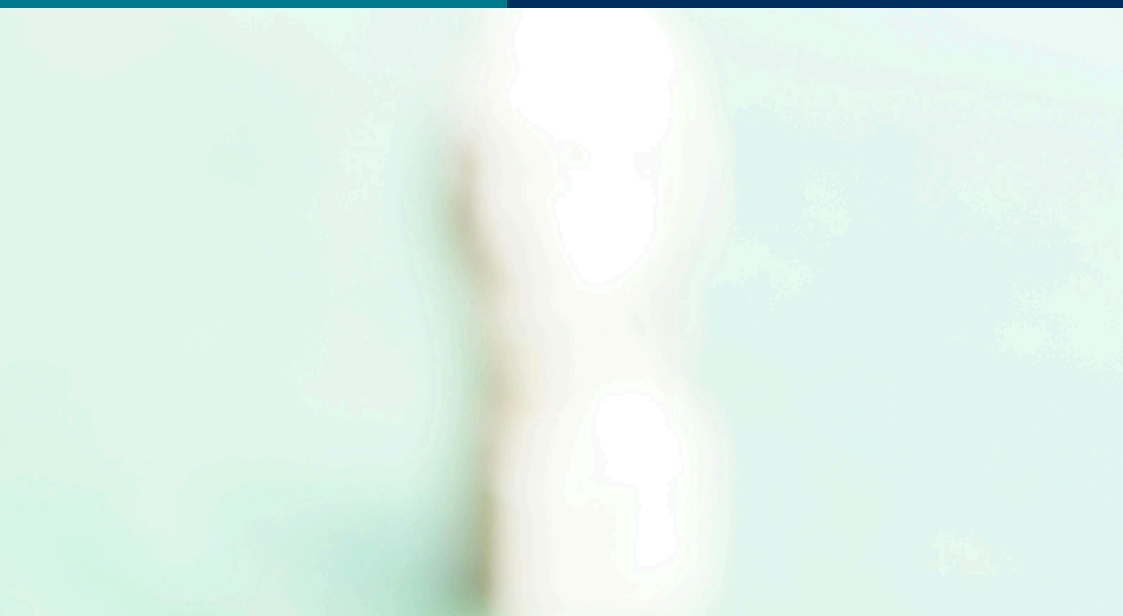
- Algorithmisches **just-in-time** Generieren der Bücher.
- „Generieren“ und „Inspizieren“ nur der „**wohlformulierten**“ Bücher.
- **Sinnhorizont** ist dabei das bereits **digital Publizierte!**



Programmieren durch Justieren

- I) **Identifikation / Konstruktion** eines für die Aufgabe passenden, hoch **parametrisierten Modells** und
- II) Einstellung von dessen Parametern über ein mathematisches **Optimierungsverfahren** so, dass es **für möglichst alle Eingaben passende Ausgaben** produziert
(**Maschinelles „Lernen“**, kurz ML).

ML-Systeme sind **algorithmisierte Assoziations-Tabellen**.



Künstliche Neuronale Netze als Standard-ML-Modell

Künstliche Neuronale Netze sind **effiziente Kompressoren** selbst **riesiger Assoziations-Tabellen...**

Aber:

- je mehr man beim Modellieren Wert auf eine starke Kompression (**möglichst kleines und effizientes Modell**) legt,
- desto häufiger werden **Kompressionsverluste** resultieren,

das trainierte System antwortet dann „unpassend“, es **konfabuliert**.

Konfabulation

- **Erfinden von Geschichten**, Ereignissen oder Informationen, die nicht der Realität entsprechen.
- Unterschiedliche auslösende Faktoren, darunter **Gedächtnislücken**, **kognitive Beeinträchtigungen** oder neurologische Störungen.
- Personen, die konfabulieren, sind sich dessen trotz klarem Bewusstseins normalerweise **nicht bewusst**.
- Das Konfabulat kann **weitgehend zusammenhanglos und auch unlogisch** sein.

Halluzination

- **Sensorische Wahrnehmung**, die **ohne einen tatsächlichen externen Reiz** auftritt: z.B. Sehen von Dingen, die **nicht existieren**, Hören von Stimmen oder Geräuschen.
- Verursacht u.a. durch **psychische Erkrankungen** wie Schizophrenie, Drogenmissbrauch, Schlafmangel oder neurologische Störungen.
- Die betroffene Person ist sich — selbst bei getrübttem Bewusstsein — normalerweise **bewusst**, dass die Wahrnehmungen nicht real sind.

„Incredible! ChatGPT can view an image and recreate it“¹:



¹(Mao 2023)



(Mao 2023)

Grundprobleme maschinell trainierter Systeme

- **Unvollständiges Wissen** über das „richtige“ Verhalten.
- **Nicht-Erklärbarkeit** des produzierten Verhaltens.
- **Bias und Verzerrungen.**



Konstruktion von ChatGPT

Zwei Hauptphasen:

- I) **Pretraining**
zur **Generierung des sog. Sprachmodells** (GPT) und
- II) **Finetuning des Sprachmodells**
für ein „natürliches“ Dialogverhalten.

ChatGPT = GPT + Finetuning

I) Pretraining

Nur orientiert an den digital vorliegenden Texten, wird hier

- für einen **beliebig vorgegebenen Diskursverlauf** und
- **jedes sprachlich theoretisch mögliche Wort**

ausgerechnet, mit welcher **Wahrscheinlichkeit** dieses Wort das **nächste**, von ChatGPT geäußerte **Wort im Diskursverlauf** sein sollte.

Das zugehörige statistische Modell nennt man sein **Sprachmodell**, seinen **Generative Pretrained Transformer (GPT)**.

Autoregressive Textassoziation mit dem Sprachmodell:

Gegeben einen Prompt des Nutzers,
ergänzt ChatGPT während eines Chats

- **entsprechend diesem Sprachmodell**
- **solange** den soweit entwickelten Chatverlauf
um je ein plausibles nächstes Wort,

bis die Folge dieser Wörter / Token aus Sicht des Sprachmodells
eine **vollständige Antwort** darstellt.

Wortassoziationstest (Markus Antonius Wirtz 2021)

„Gruppe von Verfahren, bei denen der Proband auf ein vorgegebenes **Stimuluswort** eine oder mehrere **Antworten** liefern soll, und zwar **das oder die ersten, die ihm in den Sinn kommen. ...**

In der klin. Anwendung wird heute in breiterem Maße lediglich die **fortgesetzte freie Assoziation** unter der modernen Bez. der Wortflüssigkeitstests eingesetzt, deren Ergebnisse als Indikatoren für das **semantische Gedächtnis bzw. exekutive Funktionen** gelten."

Semantisches Gedächtnis (Markus Antonius Wirtz 2021)

„bezeichnet unser generelles (Fakten-) **Wissen über die Welt, unabhängig vom eigenen Erleben** (Erfahrung). Dieses Wissen besteht aus grundlegenden Bedeutungen von Wörtern, Begriffen sowie den Beziehungen zw. ihnen.“

Exekutive Funktionen (Markus Antonius Wirtz 2021)

„Funktionen, die **kognitive Verarbeitung** steuern, bspw. den Wechsel von einer Form der Aufgabenbearbeitung zu einer anderen, die Hemmung dominanter, aber nicht aufgabenangemessener Reaktionen oder die Koordination mehrerer kogn. Funktionen bei der Mehrfachtätigkeit.“

II) Finetuning:

Iteratives Verfeinern des Sprachmodells,
durch Nutzung **aufgabenspezifischer Datensätze**, die

- passgenaue Antworten für **spezifische Fragen** enthalten oder
- aus besonders **themenbezogenen** Textmaterial bestehen.

Dieses Verfeinern erfordert einen

extrem zeit- und kostenintensiven

Einsatz von Menschen (Clickworkern):

Reinforcement Learning from Human Feedback (RLHF).

Einflussnahme auf ChatGPT

Anpassung von **Parametern**.

Vorgabe eines **für die Aufgabe optimalen Prompts**.

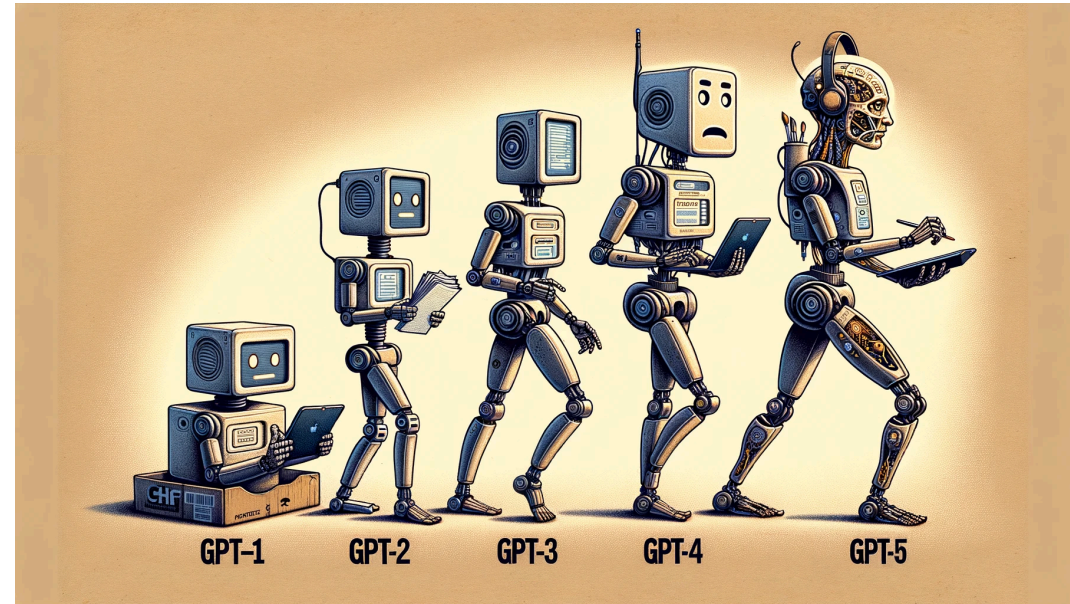
Nutzerspezifisches **Finetuning**

- für einen bestimmten Anwendungsbereich
- unter Ausnutzung der von OpenAI zur Verfügung gestellten **Programmierschnittstelle (API)** zu ChatGPT.

Ressourcen für (Chat)GPT - Wie (lange) skaliert das?

Evolution von 2017–2023:

- $1.17 \cdot 10^8 \rightarrow 1.8 \cdot 10^{12}$
Modell-Parameter,
- 23 \rightarrow 120 Zetabytes
pot. Trainingsdaten,
- 100 \rightarrow 1.100
Milliarden GigaFLOPS.



ChatGPTs Werk und des Menschen Beitrag

Wo steckt der Mensch beim Bau von ChatGPT-XY?

- I) Im **verschriftlichen kulturellen Erbe** der Menschheit und den **Massendaten**, die fortlaufend **von den Menschen abgeschnorcht** werden,
- II) im Blut und Schweiß der **Heerscharen** schlecht bezahlter **Clickworker**, die diesen Daten beim Finetuning erst Sinn geben,
- III) in den Köpfen der **Wissenschaftler und Ingenieure**, die die ChatGPT zugrundeliegende Technologie weiterentwickeln und
- IV) in den öffentlichen Institutionen, die sich durch **Vorgabe ethischer und politischer Rahmenbedingungen** ein gesellschaftskonformes **Alignment** und einen **verantwortungsvollen Einsatz** von ChatGPT erhoffen.

ChatGPT – Extrem nützlich, aber auch intelligent?

- **Ohne den Menschen ist (und wird) ChatGPT nichts.**
- ChatGPT ist ein äußerst belesener und extrem gut trainierbarer **stochastischer Papagei** (Bender u. a. 2021):

Bei Google antwortet **ein Mensch**,
bei ChatGPT die **statistisch gemainstreamte Menschheit**.

- Wer gut **sprechen** kann, versteht noch lange nicht **Sprache** oder gar das **Gesprochene**.

Quo vadis, ChatGPT?

Noch stärker **multimodale Modelle** (LLM \rightarrow LMM)

- **größere Kontextfenster** (Tokenlimit: 128k \rightarrow 1 Mio),
- Anpassung der **Finetuning-API** an die neuen, erweiterten Modelle,
- Bereitstellung einer API mit **Gesprächsgedächtnis**.
- konfigurierbare, **GPT-basierte Konfektions-ChatBOTS**

Der feuchte (**Alp?**)**Traum** der Jünger des Maschinellen Lernens:

ChatGPT++ als recursive, self improving and embedded
Artificial General Intelligence (AGI).

Münchhausen im Sumpf

„Bei der Verfolgung eines Hasen wollte ich mit meinem Pferd über einen Morast setzen. Mitten im Sprung musste ich erkennen, dass der Morast viel breiter war, als ich anfänglich eingeschätzt hatte. Schwebend in der Luft wendete ich daher wieder um, wo ich hergekommen war, um einen größeren Anlauf zu nehmen.

Gleichwohl sprang ich zum zweiten Mal noch zu kurz und fiel nicht weit vom anderen Ufer bis an den Hals in den Morast. Hier hätte ich unfehlbar umkommen müssen, wenn nicht die Stärke meines Armes mich an meinem eigenen Haarzopf, samt dem Pferd, welches ich fest zwischen meine Knie schloss, wieder herausgezogen hätte.“

Artwork: Theodor Hosemann (1807-1875), public domain



Könnte eine sprechende Maschine Mensch sein?

„Jede Maschine ist naturgemäß beschränkt in dem, was sie tun kann.

Während der menschliche Körper auch als eine Maschine betrachtet werden kann, könnte er nie durch eine künstliche Maschine nachgebildet werden.

Dies liegt daran, dass eine Maschine niemals über genügend Vielfalt verfügen würde, um als Mensch zu agieren...

*Wenn Maschinen die Fähigkeit hätten, in irgendeiner Weise zu **sprechen, wie wir es tun**, dann könnten sie einem **jeden Menschen in allen Dingen antworten**; und sie würden nicht nur in gewissen Dingen schlechter als der gemeinste Mensch agieren..."*

(gekürzt aus **Descartes'** „Abhandlungen über die Methode“, 1637)

Frage an ChatGPT:

„Ist insbesondere der zweite Teil dieses Descartes-Zitates im Zeitalter von großen Sprachmodellen (LLMs) a la ChatGPT überholt? Begründe bitte Deine Antwort.“

ChatGPTs Replik (ebenfalls leicht gekürzt):

„Obwohl LLMs wie ChatGPT enorme Fortschritte gemacht haben, um menschenähnliche Antworten in verschiedenen Kontexten zu liefern, fehlt ihnen ein tiefes Verständnis und Bewusstsein für den Inhalt ihrer Antworten.

Im Gegensatz zu Menschen, die nicht nur Worte verwenden, sondern auch Absicht, Emotionen und einen umfassenden Kontext berücksichtigen, nutzen LLMs Mustererkennung, um Verständnis zu simulieren, ohne dabei ein echtes Bewusstsein oder Emotionen zu besitzen.

*LLMs generieren Antworten auf Grundlage ihrer Trainingsdaten und **können nicht wirklich „lernen“** oder sich an Entwicklungen nach ihrem letzten Training anpassen. Zudem **fehlt ihnen die Fähigkeit zur Intuition und kreativen Problemlösung.***

*Menschliche Kommunikation bezieht moralische und ethische Überlegungen ein, die auf individuellen Wertesystemen basieren. **LLMs können Ethik und Moral lediglich auf Grundlage ihrer Trainingsdaten simulieren, besitzen jedoch kein eigenes Wertesystem.***

*Menschen können unvorhersehbar und abseits von etablierten Mustern reagieren, während **LLMs aufgrund ihrer Konstruktion deterministisch sind.***

LLMs können deshalb die Tiefen und Nuancen der menschlichen Kommunikation, die über bloße Sprache hinausgehen, noch nicht vollständig nachbilden. In dieser Hinsicht bleibt Descartes zumindest teilweise gültig."

Dank an die kreativ am Vortragsdesign Beteiligten:

- **MSc. Johannes Mitschunas** - Doktorand in der AG KI
- **Daniel Motz** - Studentische Hilfskraft in der AG KI

- **Dr. Beate Mitschunas** - Fotografie

- OpenAI für **ChatGPT und Dall-E 3**
- Stability AI für DreamStudio (**Stable Diffusion**)

Bibliographie

- Bender, Gebru, McMillan-Major, und Shmitchell. 2021. *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* 🦜. New York, NY, USA: Association for Computing Machinery
- Borges, Jorge Luis. 1992. „Die Bibliothek von Babel“. In *Fiktionen: Erzählungen 1939 - 1944*. Fischer Verlag
- Mao, Wei. 2023. „Incredible! ChatGPT can view an image and recreate it“. 7. Oktober 2023. <https://bootcamp.uxdesign.cc/incredible-chatgpt-can-view-an-image-and-recreate-it-9b243cc296b7>
- Markus Antonius Wirtz, Hrsg. 2021. *Dorsch - Lexikon der Psychologie*. 20. Aufl. hogrefe AG