Chancen & Risiken generativer KI im Unterricht und in Prüfungen

Chemnitzer Linux-Tage am 22. 3.2025
- Update 4.4.2025 -

Klaus Knopper <klaus.knopper@hs-kl.de>

Experimente und Infos zum Kurs

→ https://knopper.net/chlt/ki/

(Passwort auf Anfrage)

Künstliche Intelligenz

- ...ist nicht das, was man intuitiv damit verbindet („~~schlaue Maschinen~~“)

- *Definition* des → europ. Parlaments:
  „**Künstliche Intelligenz** ist die **Fähigkeit einer Maschine**, **menschliche Fähigkeiten** wie logisches Denken, Lernen, Planen und Kreativität zu **imitieren**.“ *)

- *Definition* nach Stuart J. Russell und Peter Norvig**) in vier unabhängigen Kategorien:

  1) **Nachbildung** von **menschlichem Denken**
  2) **Nachbildung** von **rationalem Denken**
  3) **Nachbildung** von **menschlichem Verhalten**
  4) **Nachbildung** von **rationalem Verhalten**

Während 1. und 2. eher eine „**Simulation menschlichen Verhaltens**“ als Ziel definieren (vergl. frühe erfolgreiche Versuche mit **„Eliza“ von Joseph Weizenbaum 1966*****), fließt in 2. und 3. auch eine menschenunabhängige Art **„Vernunft“-Simulation** ein.

*) Was ist künstliche Intelligenz und wie wird sie genutzt? | Aktuelles | Europäisches Parlament. 14. September 2020, abgerufen am 18. Juli 2021.
**) Stuart J. Russell, Peter Norvig: Künstliche Intelligenz: Ein moderner Ansatz. Pearson Studium, Berkeley 2004, ISBN 3-8273-7089-2 (englisch: Artificial Intelligence: A Modern Approach.).
***) → https://de.wikipedia.org/wiki/ELIZA

Was bedeutet „generative KI“ eigentlich?

- Inhalte **erzeugen** (Input → Output, auch „Transformation“)
- I.d.R. mit vortrainiertem („pre-trained“) großem Sprachmodell („**L**arge **L**anguage **M**odel“)
- GPT: **G**enerative **P**re-trained **T**ransformer
- Funktionsweise, eigener Definitionsversuch: **Berechnung** von **Wahrscheinlichkeiten** für auf **eine Frage** „am **besten passende**“ **Text/Grafik/Audio-Bausteine** („Tokens“) – mit Steuerungsmöglichkeit – nebst **schrittweisem Zusammenfügen einer „optimalen“ Antwort** mit einstellbarer Varianz („Zufall“), **aber ohne „semantisches Verständnis“ des generierten Inhalts** durch den Algorithmus.

→ **Live-Demo auf Raspberry Pi** starten, dauert etwas…

KI-Tools sind toll...

- **Erzeugen** von **professionell bis „perfekt“ gestalteten**, **zielgruppenspezifischen Inhalten**.

- **Texte**, **Bilder**, **Sprache**, **Musik**, **Programmcode**...

- Im **Browser**, per **App**, mit **Sprach**ein-/ausgabe **nutzbar**.

- KI-Tools **großer Firmen**: **Niederschwellig**, einfach zu bedienen (gelegentlich sogar **barrierearm**), **umfangreiche Datenbasis**, mit oder ohne **Anbindung an Internet-Recherche**.
- KI-Tools können **Lösungsansätze vorschlagen**, wenn man **selbst keine Idee hat**, wie eine Aufgabe zu lösen ist.
- KI-Tools können **sehr schnell Lehrmaterial generieren** (nein, diese Folien sind nicht von ChatGPT erstellt. ;-)

ChatGPT, DeepL & Co. vs. offene KI-Tools (Deepseek, Gemma, Meta-Llama, ...)

- **Bei vielen Sprachmodellen sind die Gewichte als Ergebnisse aufwändiger Trainings die “Kronjuwelen”, mit denen sich Dienste und Software-Produkte verkaufen / vermieten lassen. → Proprietäres Geschäftsmodell.**

- **Inzwischen gibt es viele “offene” (wie offen, schauen wir uns gleich an) Modelle und Open Source Engines.**
  Demo: Beispiel → ollama

- **Universeller Anspruch größerer Modelle vs. Spezifischer Einsatz von Destillaten / quantifizierten Modellen. → später.**

- **Kurzer Exkurs (aus aktuellem Anlass) zu Deepseek. → jetzt**

Über Deepseek (Unternehmen)

DeepSeek ist ein **chinesisches Startup**, das sich auf die **Entwicklung fortschrittlicher Sprachmodelle und künstlicher Intelligenz spezialisiert** hat. Das Unternehmen gewann internationale Aufmerksamkeit mit der Veröffentlichung seines im **Januar 2025 vorgestellten Modells DeepSeek R1**, das mit **etablierten KI-Systemen wie ChatGPT von OpenAI und Claude von Anthropic konkurriert**. Das Unternehmen wird **ausschließlich vom chinesischen Hedgefonds High-Flyer** finanziert. Beide Unternehmen haben ihren Sitz in **Hangzhou, Zhejiang**. *)

*) Quelle: → Wikipedia 31.1.2025

Hangzhou DeepSeek Artificial Intelligence Basic Technology Research Co., Ltd.

deepseek

Rechtsform	Limited Company
Gründung	März 2023
Sitz	Hangzhou, Volksrepublik China
Leitung	Liang Wenfeng (CEO)
Branche	Informationstechnik Künstliche Intelligenz
Website	www.deepseek.com

Stand:

Über Deepseek (Software)

Am **20. Januar 2025** präsentierte **DeepSeek** das Large Language Model **DeepSeek-R1**, das auf **maschinellen Lerntechnologien** basiert und eine Architektur verwendet, die konzeptionell mit den gängigen Transformer-Modellen vergleichbar ist.

DeepSeek-R1 wurde unter der **MIT-Lizenz** veröffentlicht, die **uneingeschränkten Open Access fördert und kommerzielle als auch akademische Nutzungen ohne Einschränkungen erlaubt**. Damit setzt das Unternehmen bewusst einen **Kontrast zu zahlreichen proprietären KI-Systemen**, die durch restriktive Lizenzen gekennzeichnet sind.

Der **Zugang** zu DeepSeek ist möglich durch
- die DeepSeek-**App**,
- die **Webseite**,
- die **Programmierschnittstelle (API)** oder
- die I**nstallation des Anwenders auf einem eigenen Computer**.

DeepSeek senkte laut dem Einzelunternehmer Cleanthinking.de den **Energieverbrauch im Vergleich zu traditionellen KI-Modellen um bis zu 70 %** durch **effizientere Algorithmen**. Allerdings gibt es **Zweifel** daran, ob DeepSeek die **Modelle vollständig selbst trainiert hat**, oder ein bereits fertig trainiertes **ChatGPT als Hilfsmittel** nutzte, um die Trainingszeit zu verkürzen.

Im Gegensatz zu bisherigen Modellen mit überwachter Feinabstimmung (SFT) nutzt DeepSeek **R1** das auf Millionen von Inferenzspuren trainierte **Reinforcement Learning (RL)**. Dies **imitiert menschenähnliche Bewertungen**, was eine tiefere Analyse von Aufgaben ermöglicht, die komplexe Schlussfolgerungen erfordern. DeepSeek **verwendet „Aha-Momente“** als Pivot-Token bei der **Formung von Schlussfolgerungen (Chain-of-Thought, CoT)**. Diese „Aha-Momente“ dienen dazu, **Zwischenschritte zu reflektieren und neu zu bewerten**. Damit wird die **Qualität der Antworten durch Selbstkorrektur verbessert**. Veröffentlicht wurden für **rohe Schlussfolgerungen** die Variante **DeepSeek-R1-Zero** sowie **DeepSeek-R1 für praktische Anwendungen**.

*) Quelle: → Wikipedia 31.1.2025

Deepseek-API

Die **Programmierschnittstelle (API)** von Deepseek ist **kompatibel mit der veröffentlichten API von OpenAI**, daher sind minimale Änderungen notwendig, um auf Deepseek zu wechseln.

Beispiel in PHP:

```
    $question = 'Löse die Gleichung x²+y³=100';
    $apiKey = loadApiKey('/path/to/API_KEY_DEEPSEEK');
//  $url = 'https://api.openai.com/v1/chat/completions';
    $url = 'https://api.deepseek.com/chat/completions';
    $data = [
//      'model' => 'gpt-4o-mini',
        'model' => 'deepseek-chat',
        'messages' => [['role' => 'user', 'content' => $question]],
        'max_tokens' => 1600
    ];
    $options = [ 'http' => [  'header'  => "Content-type: application/json\r\n" . "Authorization: Bearer $apiKey\r\n",
            'method'  => 'POST', 'content' => json_encode($data) ] ];
    $context  = stream_context_create($options);
    $result = file_get_contents($url, false, $context);
```

Wie bei OpenAI ist eine kostenpflichtige Registrierung zur Nutzung der API bei Deepseek notwendig, um ein API-Token zur Autorisierung und Abrechnung zu erhalten.

Über Janus (Software)

Mit **Janus-Pro** hat DeepSeek am 27. Januar 2025 ein neues **multimodales KI-Modell** veröffentlicht, das unter der **MIT-Lizenz als Open Source** verfügbar ist. Das Modell kombiniert **Fähigkeiten**, die sowohl **DALL-E** als auch **Stable Diffusion** ähneln, und kann **Bilder erzeugen sowie analysieren**. *)

*) → Wikipedia 2.2.2025

„a cute cat“

„sunset in the rhineland-palatine forest“

„a fantasy-style picture of buildings of the university of applied sciences kaiserslautern“

Ist Deepseek (Software) Open Source?

Engine/Code (Python): MIT Lizenz
Daten (Gewichte/Parameter): MIT Lizenz (geändert 20.1.2025 → Changelog)

Gemäß der **AI Open Source Definition** von → https://opensource.org/ai/open-source-ai-definition:

Code: The complete source code used to train and run the system. The Code shall represent the full specification of how the data was processed and filtered, and how the training was done. Code shall be made available under OSI-approved licenses. ✓

Parameters: The model parameters, such as weights or other configuration settings. Parameters shall be made available under OSI-approved terms. ✓

Data Information: Sufficiently detailed information about the data used to train the system so that a skilled person can build a substantially equivalent system. Data Information shall be made available under OSI-approved terms. ✗

Deepseek als Ganzes ist **nicht Open Source**, auch wenn die Entwickler alle **veröffentlichten Quellen unter MIT Lizenz** gestellt haben, da der **Zugriff auf die Rohdaten** für das Training **nicht möglich** ist und die Parameter / Gewichte daher **nicht reproduziert werden können**.

Was ist neu an Deepseek R1?

„Reasoning Model“: Der Prozess, mit dem eine Antwort entsteht, lässt sich über Zwischenschritte beobachten und ggf. optimieren. Beispiel:
Frage: Löse die Gleichung $x^2 + y^3 = 100$
Antwort:

Okay, ich muss die Gleichung $x^2 + y^3 = 100$ lösen. Hmm, was genau bedeutet das? Soll ich alle ganzzahligen Lösungen finden, oder reelle Lösungen? Die Aufgabenstellung sagt nicht explizit, welche Art von Lösungen gesucht sind. Normalerweise, wenn in solchen Gleichungen nach Lösungen gefragt wird, ohne weitere Angaben, könnte es sich um ganzzahlige Lösungen handeln. Aber ich sollte vielleicht beide Möglichkeiten überlegen.

Zuerst denke ich an ganzzahlige Lösungen. Also x und y sollen ganze Zahlen sein, die die Gleichung erfüllen. Das heißt, x^2 ist ein Quadrat einer ganzen Zahl und y^3 ein Kubik einer ganzen Zahl, und ihre Summe ergibt 100.

Um das zu lösen, könnte ich versuchen, mögliche Werte für y zu finden, so dass $y^3 \leq 100$, denn x^2 muss ja dann $100 - y^3$ sein, und das muss nicht-negativ sein, da x^2 immer ≥ 0 ist.

Also, $y^3 \leq 100$. Welche ganzen Zahlen y erfüllen das?

y=0: $0^3=0$, okay

y=1: $1^3=1$

y=2: 8

y=3: 27

y=4: 64

y=5: 125, das ist schon größer als 100. Also y kann maximal 4 sein.

Und negative y? y=-1: $(-1)^3=-1$, dann $x^2 = 100 - (-1) = 101$. 101 ist keine Quadratzahl, denn $10^2=100$ und $11^2=121$. Also $x^2=101$ hätte keine ganzzahlige Lösung. y=-2: $(-2)^3=-8$, $x^2=100 - (-8)=108$. 108 ist auch keine Quadratzahl. y=-3: $(-3)^3=-27$, $x^2=100 - (-27)=127$. Auch keine Quadratzahl. y=-4: $(-4)^3=-64$, $x^2=100 - (-64)=164$. Keine Quadratzahl. y=-5: $(-5)^3=-125$, $x^2=100 - (-125)=225$. 225 ist 15^2. Oh, das ist interessant. Also y=-5, dann $x^2=225$, also x=15 oder x=-15. Das wären also Lösungen. Aber y=-5 ist ein möglicher Wert. Also, die möglichen y-Werte für ganzzahlige Lösungen sind y=0,1,2,3,4,-5.

Jetzt prüfe ich jeden dieser y-Werte, ob $100 - y^3$ eine Quadratzahl ist.

y=0: 100 - 0 = 100. 100 ist 10^2, also x=±10. Also (10,0), (-10,0).

y=1: 100 -1 =99. 99 ist keine Quadratzahl. $9^2=81$, $10^2=100$, dazwischen liegt 99, also keine Lösung.

y=2: 100 -8=92. 92 ist keine Quadratzahl. $9^2=81$, $10^2=100$, 92 dazwischen.

y=3: 100 -27=73. 73 ist keine Quadratzahl. $8^2=64$, $9^2=81$.

y=4: 100 -64=36. 36 ist 6^2. Also x=±6. Also (6,4), (-6,4).

y=-5: 100 - (-125)=225, wie oben erwähnt. x=±15. Also (15,-5), (-15,-5).

Gibt es noch andere negative y? y=-6: $(-6)^3=-216$, dann $x^2=100 - (-216)=316$. 316 ist keine Quadratzahl ($17^2=289$, $18^2=324$). Also keine Lösung.

Also insgesamt die ganzzahligen Lösungen:

(±10, 0), (±6,4), (±15,-5)

Kann man Deepseek R1 auch selbst betreiben (ggf. ohne Internet-Anbindung)?

- **Grundsätzlich ja**, und darf dies aufgrund der Lizenz auch sowohl privat als auch kommerziell.

- Hardware-Mindest-Anforderungen, um das vollständige R1-Modell sinnvoll zu betreiben:
  RAM/GPU-Bedarf
  671 Milliarden (Parameter) * 16bit (int od. floating point) / 8 (bit pro byte)
  = 1342000000000 Bytes
  = **1,342 TB RAM**
  (z.B. zehn →H200-GPUs mit jeweils 141 GB) *) **) **)

*) Quelle: https://www.heise.de/hintergrund/DeepSeek-Blick-hinter-die-Kulissen-des-Reasoning-Modells-R1-10260703.html
**) Preis pro Stück 35000€, Stand 2.2.2025
***) Mit schichtenweiser Reduktion der Genauigkeit (int/float bits), auch mit >= 20GB und CPU möglich, aber „kartoffelig“ langsam, s.a. https://unsloth.ai/blog/deepseekr1-dynamic

Was sind „Destillate“?

Eine „**kleinere Version**“ der großen Deepseek V2 oder R1 Datenbank ist, für bestimmte Anwendungen optimiert (z.B. Programmcode-Generierung, aber auch Chat), als sog. **Destillat** verfügbar. Dabei wird ein **neues Modell mit weniger Parametern** erzeugt, das die umfangreicheren Datensätze des großen Modells mit einem (deutlich kleineren) **Basismodell** durch Nachtrainieren „**filtert**“, so dass immer noch „akzeptable“ Antworten trotz **deutlich geringeren Datenmengen und Ressourcenanforderungen** entstehen.

Model	Base
DeepSeek-R1-Distill-Qwen-1.5B	Qwen2.5-Math-1.5B
DeepSeek-R1-Distill-Qwen-7B	Qwen2.5-Math-7B
DeepSeek-R1-Distill-Llama-8B	Llama-3.1-8B
DeepSeek-R1-Distill-Qwen-14B	Qwen2.5-14B
DeepSeek-R1-Distill-Qwen-32B	Qwen2.5-32B
DeepSeek-R1-Distill-Llama-70B	Llama-3.3-70B-Instruct

Es ist bei Deepseek durch die Veröffentlichung der großen Modelle auch **möglich**, **eigene Destillate** zu erstellen. V.a. mit dem R1-Modell als Basis lässt sich dabei nachvollziehen, ab welcher Komplexität das destillierte Modell bei einer Fragestellung scheitert bzw. falsche Antworten liefert.

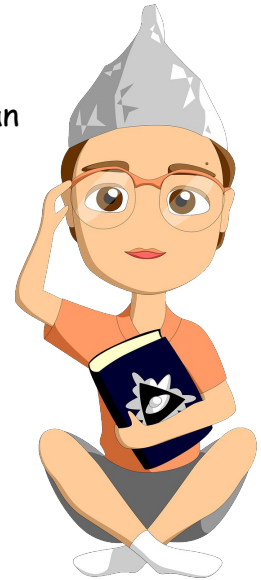
Fake Information oder Zensur (1) ?

- Das veröffentlichte R1-Datenbankmodell: Faktenanteil unklar, zumindest in bisherigen Tests kaum oder wenig offensichtliche Fehler, vergleichbar mit ChatGPT.
- Aber: Offensichtliche Zensur bei Zugriff über Deepseek.com Webseite, Beispiel!

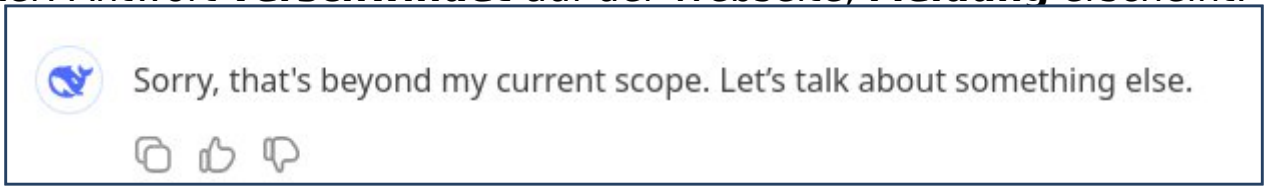
Neutrale Frage (Webseite): Wo liegt Taiwan?

Antwort beginnt:

Taiwan liegt vor der Südostküste Chinas und ist die größte Insel der Republik China (ROC). Geografisch befindet es sich im westlichen Pazifischen Ozean, getrennt vom chinesischen Festland durch die Taiwanstraße. Politisch ist Taiwan ein umstrittenes Gebiet, da die Volksrepublik China (VR China) Anspruch auf die Insel erhebt und sie als eine ihrer Provinzen betrachtet, während Taiwan sich selbst als unabhängiges Land mit einer eigenen Regierung und Verfassung sieht. Die internationale Gemeinschaft ist in dieser Frage gespalten, wobei einige Länder Taiwan als unabhängigen Staat anerkennen und andere die Position der VR China unterstützen.



2 Sekunden später: Antwort **verschwindet** auf der Webseite, **Meldung** erscheint:



Fake Information oder Zensur (2) ?

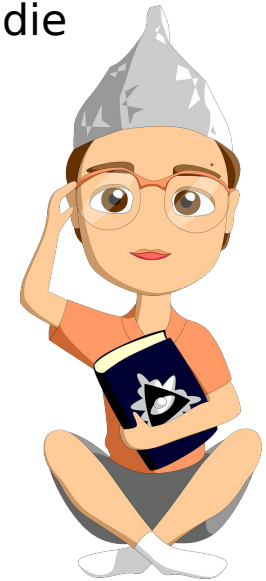
Mögliche Interpretation:

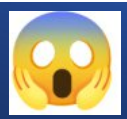
Deepseek (Unternehmen) setzt die gesetzlichen Vorgaben Chinas um, allerdings technisch etwas nachlässig (oder absichtlich offensichtlich?).

Ob und wie die veröffentlichten Modelldaten selbst politisch tendenziell oder postfaktisch trainiert wurden, lässt sich derzeit nicht sicher belegen (ist auch schwierig, ohne die verwendeten Rohdaten und Bewertungsschemata zu kennen).

Lokal gestartet per ollama tritt der “Zensur-Effekt” nicht auf.

Update: Seit ca. 1.4.2025 gibt auch die API über Deepseek-China “politisch korrigierte” Antworten. → Verlässlichkeit der API in Frage gestellt.



Nutzungsbedingungen und Datenschutz  

Mit der Nutzung der Deepseek-Dienste werden die Nutzungsbedingungen und Privatspären-Regelungen akzeptiert, die (für europäische Verhältnisse) **ähnlich verstörend** sind wie die anderer bekannter KI-Dienste und Cloud-Software – allerdings (zusätzlich) auch nach chinesischer Jurisdiktion.

Dabei ist Deepseek vergleichsweise klar und deutlich bei der Angabe der Datenerfassung und Erwartungen bei der Nutzung seiner selbst angebotenen Cloud-Dienste.

Einige Highlights:

→ DeepSeek-Terms-of-Use

- Keine Inhalte, die [auch chinesischen] Restriktionen unterliegen. (3.1ff)
- **Sie** (als Nutzer*in) sind **alleine verantwortlich** für die **Einhaltung von Deepseeks Konditionen** und **juristischen Implikationen**, und zu **Schadenersatz** und **Kostenübernahme für juristische Auseinandersetzungen** zuständig. (7.2)

→ DeepSeek-Privacy-Policy

Zusammengefasst: **Alle** auf **irgendeine Weise** erfassbaren **Informationen**, Eingaben, automatisch erfassbare technischen Daten, z.B. Betriebssystem, verwendete Software, genutzte Dienste Dritter wie **Social Media Plattformen und dort gespeicherte Informationen** bis hin zum **Nutzerverhalten bei Tastaturbedienung (Timings)** und Nutzungsprofilen in Webshops u.ä. **werden von Deepseek erfasst, gespeichert, verarbeitet und mit Partnerunternehmen sowie „law enforcement“ und „government“ geteilt**.

Chancen (speziell bei Deepseek)

- Die **Veröffentlichung (teilweise) unter einer Open Source-Lizenz** erlaubt es, die verwendeten Algorithmen beim Betrieb des Modells besser zu verstehen und sich an der Weiterentwicklung zu beteiligen.

- Besonderheit der **MIT-Lizenz**: Es ist **erlaubt**, auch **proprietäre Software (ohne Offenlegung des Quellcodes**) unter **Verwendung der veröffentlichten Algorithmen und Daten zu schreiben**.

- **Kosten** und **Ressourcenverbrauch** sind, verglichen mit den bisherigen „Abo-Modellen“ proprietärer Anbieter, deutlich niedriger, auch dann, wenn man das Mietmodell von Deepseek (kostenpflichtige Nutzung der Deepseek-Server und/oder API) wählt.

- Neben dem Chat sind auch **Bildgenerierung** (→ Janus) und KI-gestützte **Programmierung** (→ deepseek-coder) möglich und werden ebenfalls sowohl als Open Source, als auch im Mietmodell als Webdienst angeboten.

Risiken

- Zugriff über **Deepseeks eigene Web-Dienste und die offizielle Deepseek-Web-API zensiert/manipuliert Antworten** (sehr offensichtlich).
- **Herkunft der Rohdaten** fürs Training und ggf. **Einfluss auf Gewichte unklar** (das ist allerdings auch bei anderen GPT-Modellen so).
- **Kein Datenschutzabkommen** mit China, alle Daten, die über die Deepseek-API/Webseite geschickt werden, könnten prinzipiell für alle Zwecke verwendet werden.
- **Zukünftige Versionen** der Engine und/oder **Daten** könnten eine **andere Lizenz** erhalten, die jetzige sehr **offene Veröffentlichungspraxis** ist **für die Zukunft nicht garantiert**.

Zurück zum übergeordneten Thema...

3 Shades of Digitale Souveränität in KI

1. “Ich kann mir das **KI-System** mit den **jeweiligen Lizenzbedingungen selbst aussuchen** (fast immer)”
   → geht auch proprietär / Miete

2. Wie 1., und “ich **kann und darf** das **KI-System jederzeit**, **ohne Einschränkungen**, **auf eigener Hardware oder bei einem Dienstleister meiner Wahl nutzen**”
   → Kombination Open Source Engine plus Open Weights Sprachmodell

3. Wie 2., und “ich **kann und darf** den **Aufbau** und die **Architektur des KI-Systems nachvollziehen** und **uneingeschränkt verändern**.
   → Open Source für beides, Engine und Sprachmodell

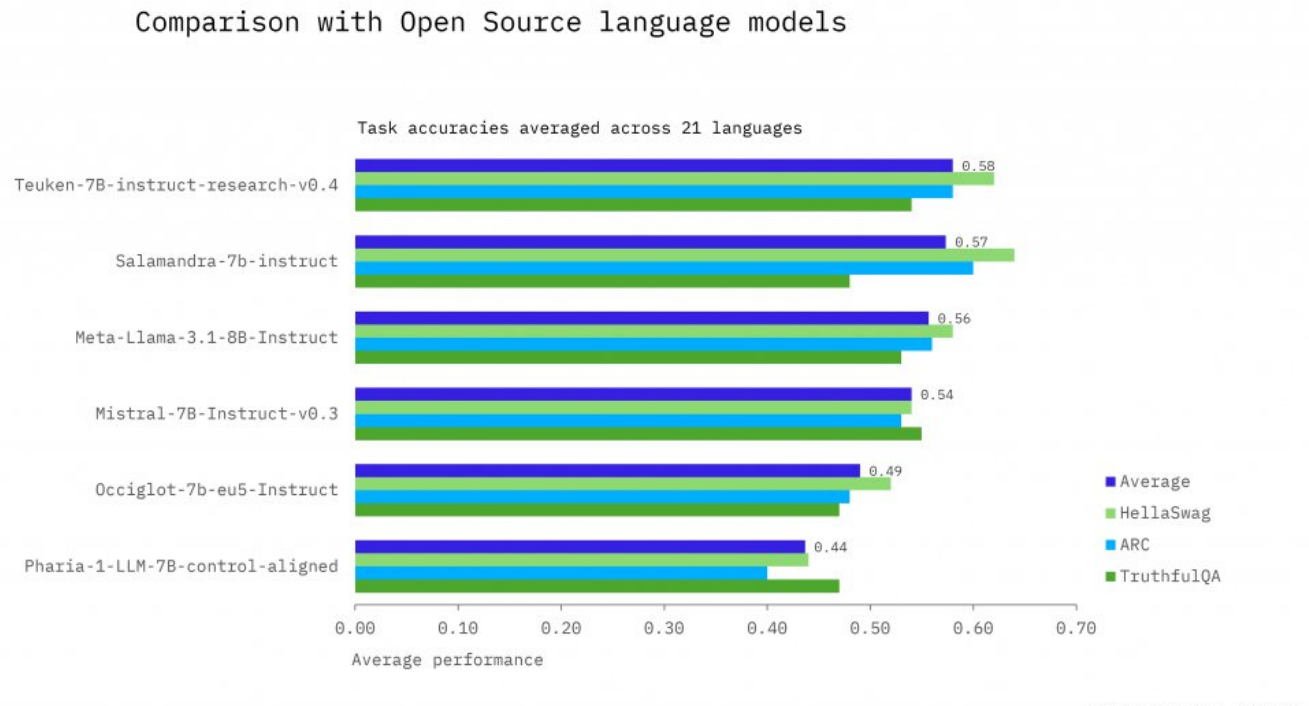
§§§

Europäisches Sprachmodell-Projekt (Open Source, auch Trainingsdaten)

→ OpenGPT-X: Trainingsdaten in 24 europ. Amtssprachen, "research" und "commercial-grade" Modell unter Open Source Lizenz (Fraunhofer Institut + DFKI)

Veröffentlicht: → Teuken-7B (7 Milliarden Parameter)

Modelle und Trainingsdaten auf → Huggingface



Mindset Digitale Souveränität in KI - KI Kompetenz

- Beim **Arbeiten mit KI** die generierten Resultate **nicht ohne Prüfung und eigenes Hintergrundwissen** für weitere Bearbeitungen heranziehen.
- **Vielfalt von Sprachmodellen** für **verschiedene Zwecke** ausprobieren.
- **Verstehen**, wie die **Generierung funktioniert**.
- **KI-Verordnung**: **→ Kompetenzpflicht/-nachweis ab 2.2.2025 zur dienstlichen Nutzung von KI-Tools in Europa**.

Die Philosophie-Seite: Weiß/versteht der KI-Algorithmus, was er tut?

- **Falsche Frage.** ;-)
  Wir sehen live, dass eigentlich nur Wortschnipsel aufgrund einer Heuristik aneinandergereiht werden.

- Je **ähnlicher** der Output einer **menschlichen Antwort** kommt, desto mehr hat man allerdings **den Eindruck**, dass die Maschine „denkt“ oder „versteht“
  → **bessere Simulation**.

- Auch die Frage, ob eine **KI ein „Bewusstsein“ entwickeln** könnte, ist akademisch (zumindest, so lange wir noch nicht einmal richtig verstehen, **wie das menschliche Gehirn funktioniert**).

- Wenn es hilft/glücklich macht, sich **vorzustellen**, die KI würde den Menschen **persönlich verstehen** und könnte sogar ein*e „elektronische*r Freund/Freundin“ sein → *go for it*, aber **Awareness für die technischen Zusammenhänge ist wichtig, um Enttäuschungen und Gefahren zu vermeiden**.

Was ist denn das Problem? → Herausforderungen

- **Kommerzielles Interesse** der KI-Anbieter liegt **nicht in der Präzision / universellen Einsetzbarkeit,** „Wahrheit“, sondern eher „**Gefallen**“ (**Kundenbindung erwünscht**)
  → **Faktencheck wichtiger als je zuvor**!

- Ohne eigenes Hintergrundwissen bzw. „begleitetes Surfen“ **für Nutzer schwierig**, zwischen „**Fake Information**“ und **Fakten zu unterscheiden**. Auch „**Halluzinationen**“ der KIs möglich und bei **unzureichendem eigenen Wissen schwer zu erkennen.**
  → **Kern-Aufgabe der Lehrer*innen kann systembedingt nicht von einer KI übernommen werden** (was in gewisser Weise eine beruhigende Art Jobsicherheit für uns ist, s.a. → Job-Futuromat)

- **Urheber*innenschaft** der Rohdaten i.d.R. unklar → Problem für Zitation

- **Eigene Leistung der Schüler*innen fair bewerten**?

- **Bias** → Art der **Fragestellung beeinflusst Ergebnis** („stets gefällige Antworten“)

Beispiel Chat-Bot im Lernmanagementsystem

- **Chatbots** sind **hilfreich**, damit Schüler*innen/Studierende sich **rund um die Uhr** ooder bei Bedarf **beraten lassen können**, ohne sich sorgen zu müssen, dass der*die Lehrer*in “überstrapaziert”, wird.
- Chatbot kann Dinge **mehrfach** und unter **verschiedenen Blickwinkeln erklären** (auch **unterschiedliche Sprachmodelle** wählbar, s.a. → Edu-KI-Chat (HAWKI) Beispiel).
- **Ohne** sich – ggf. auch mit Hilfe von Chatbots – weiteres **Hintergrundwissen** anzueignen, werden auch richtige **Antworten aber teilweise nicht verstanden**:
  “Der Chatbot hat mir eine Lösung geschrieben, die nicht funktioniert!”
- **Kompetenzerwerb** zu Beginn jeder Lehrveranstaltung im **Umgang** mit den **jeweils zur Verfügung gestellten KIs erforderlich**!
- **Hinweis** darauf, dass die **Antwort** auch **trotz überzeugender Darstellung** komplett **falsch sein kann**, wenn der **Chatbot die Frage nicht richtig versteht** oder schlichtweg **keine ausreichenden Informationen** besitzt (**z.B. Modell zu klein**).

Juristisches

- **DSGVO** (hatten wir schon):
  Nutzung von KI-Tools von außereuropäischen Anbietern problematisch, auch wenn Rechenzentren im EU-Raum, da Inhalte ausgeleitet werden.
  Integration von entsprechenden „Assistenten“ in Standardsoftware muss ggf. vor der Nutzung deaktiviert werden.

- **KI-Verordnung** (hatte wir auch schon):
  → Amtsblatt der europäischen Union
  Risikobasierter Ansatz; Definition von sog. „**KI-Hochrisiko-Systemen**“ (→ Anhang III), für welche **erhöhte Anforderungen** und **Dokumentations-/Kennzeichnungspflicht** gelten, die aufgelisteten Hochrisiko-Bereiche enthalten u.a. auch Einsatz von KI in der Bildung, v.a. bei **KI-gestützter Bewertung von Leistungen** sowie **Zugangsprüfungen** und **Bewerber*innenauswahl**.

- **Urheberrecht** → S. Nächste Folien.

Ist ein KI-generierter Text ein Plagiat?

- Das **Urheberrecht** ist ein sog. **natürliches Recht**, das **ausschließlich Menschen** zusteht und auch ohne „Vertrag“ alleine **durch Schaffung eines Werks** entsteht bzw. anwendbar ist.
- Daher **können KIs keine Urheber** sein.
- Daher handelt es sich bei einem **KI-generierten Text / Bild etc. NIEMALS um ein Plagiat** („Aneignung von Leistungen anderer Menschen“), und die **„Plagiats-Kriterien“ greifen** (zumindest für die Anwender) **nicht**.
- **Menschengemachte Rohdaten**, die KIs als Basis verwenden, können sehr wohl **urheberrechtlich geschützt** sein. Da sie aber nur in „Schnipseln“ verwendet werden ist die **Signifikanz einzelner Werke im Output schier unmöglich zu klären**.
- **Unwissenschaftliche Zitation** bzw. **Verwendung von Elementen ohne Quellen-Nachweis** kann dennoch zur **Abwertung bis zum Nichtbestehen der Prüfung** führen.
- **Regeln** für **KI-generierte Inhalte** notwendig.

Kann man KI-generierte Inhalte, ggf. durch KI -Tools, erkennen?

- Die kurze Antwort: **Nein**, inzwischen (seit GPT-4*) wirklich nicht mehr.
- **Heuristische Ansätze** (Vergleich von KI-generierten Texten mit gleichem Thema mit der abgegebenen Arbeit, z.B. GptZero) **funktionieren nicht mehr**.
- **Paradox**: Durch **KI modifizierte Texte** werden u.U. von **KI-Checkern eher als menschlich bewertet**, als **menschengenerierte Originale**, neues Geschäftsmodell: „KI-based humanifier“ *)
- Da zuverlässiger Täuschungsnachweis bei KI-Verwendung unmöglich
  → **Verbote wirken nicht**.
- Lösungsansatz 1 (Prager Wirtschaftsuni): **Abschaffung** der **schriftlichen Abschlussarbeiten** (?) **)
- Lösungsansatz 2: Ausgewählte (oder alle?) **KI-Tools erlauben** und **Regeln festlegen** plus **Aufwertung eines Pflicht-Kolloquiums zur schriftlichen Arbeit**.

*) → https://www.bloomberg.com/news/features/2024-10-18/do-ai-detectors-work-students-face-false-cheating-accusations
**) → https://www.zdf.de/nachrichten/politik/ausland/chatgpt-kuenstliche-intelligenz-bachelorarbeit-100.html

Unser Ansatz zur Regulierung (HS-KL)

- **Art** und **Umfang** von **KI-generierten** oder **KI-modifizierten** Inhalten zu **erlauben oder auch nicht**, **obliegt den Prüfenden**.
- Die **Prüflinge** erhalten **vor jeder Prüfung von den Prüfenden** die **hierfür geltenden Vorgaben**, als → verbindliche Handreichung.
- **Beispiele** zu **sinnvollen Zitationsformaten** bei KI-Elementen (z.B. **verwendeter Prompt**, **Version der verwendeten Software**, **vollständige Inputs/Outputs im elektronischen Anhang**).
- **Ziel: Nachvollziehbarkeit / Plausibilität für die Prüfenden herstellen.**
- Entbindet die **Prüflinge** nicht vom **Fakten-Check** (→ ChatGPT ist keine Primärquelle)
- Entbindet die **Prüfenden** nicht vom **Fakten-Check** (→ Nachvollziehbarkeit der von der KI mitverwendeten Primärquellen).
- **Wenn nicht nachprüfbar**, sind **KI-generierte Abschnitte in Frage zu stellen** (→ unwissenschaftliche Zitation).

Schlimme Fehler

- Einen **Chatbot nach seiner Meinung fragen**.
- „**Bias**“ bereits **in der Frage** („Erkläre mir, warum … schlecht / gut ist“) oder im System-Prompt (“Du bist ein Lügner und Verschwörungstheoretiker”)
  → Demo mit NiceGPT und EvilGPT
- **Chatbot** wie eine „**Suchmaschine**“ benutzen (Ergebnisse sind immer generiert, nicht zitiert, **außer** teilw. bei RAGs)
- Aktuellen **Kontext** / Rolle bei **Chatbots** nicht kennen
- …

Neue Aufgaben

- **Digitalkompetenz stärken** (auch Verständnis fördern, wie KI-Tools arbeiten, was sie können, was nicht). → Nutzer*innen **nicht** nur **Konsumenten**, sondern **aktive Mitgestalter**.

- **Kritische Auseinandersetzung** mit **KI-generierten Inhalten** fördern (→ s.a. „EvilGPT“-Chatbot)

- **Datenschutz-Awareness** („Woher stammen die Daten, wohin gehen unsere Eingaben?“ → wichtig für Forschungsfragen → Gültigkeit von transatlatischen Datenschutzabkommen jetzt und zuküftig fraglich → US „Patriot Act“, „Cloud Act“ ff.)

- **Open Source LLMs** für spezielle Aufgaben, z.B. eigene Inhalte der (Hoch-)Schule und Lehrmaterialien einbinden („Erklär mir mit einfachen Worten … und zeige mir, wo es in welchem Dokument steht.“, s.a. FAIRD-Projekt der HS-KL) *)

*) Hier geht es NICHT nur um den Kosten-Aspekt, sondern um digitale Souveränität und eigene Gestaltungsmöglichkeiten.

Bonus-Folie: KI auf dem Raspberry Pi 5 (Vorbereitung)

- **Schnelle, ausreichend große SSD zum Laden der KI-Modelle**

- **RAM, the final frontier:**

```
dd if=/dev/zero of=/swapfile bs=1M count=20000
mkswap /swapfile
```

→ /etc/rc.local:

```
#!/bin/bash
modprobe zram
echo 10000000000 > /sys/block/zram0/disksize
mkswap /dev/zram0
swapon -p 1 /dev/zram0
swapon /swapfile
```

Hinweis: zram ist eine komprimierende Ramdisk, die bei gut komprimierbaren Daten (Nullen, Texte, ...) das Auslagern auf Flash oder Disk teilweise unnötig macht / reduziert (wird in Knoppix auch verwendet).

Bonus-Folie: KI auf dem Raspberry Pi 5 (ollama)

Unsicher, am Softwaremanagement vorbei, aber geht am schnellsten:

```
curl -fsSL https://ollama.com/install.sh | sh
```

Bereits installierte Sprachmodelle auflisten:

```
ollama list
```

Sprachmodelle suchen unter https://ollama.com/search

Sprachmodell installieren/Updaten und starten, z.B.:

```
ollama run deepseek-coder-v2:16b
```

Hilfe mit /?

Weiteres in der ollama-Shell:

```
/show info
/show parameters
/set parameter ctx_num 100000
...
```

ENDE

Dieser Vortrag steht unter einer Creative Commons Lizenz

http://creativecommons.org/licenses/by-nd/4.0/

Klaus.Knopper@hs-kl.de