

Warum lügen KIs?

und warum und wie sie trotzdem nützlich sind.

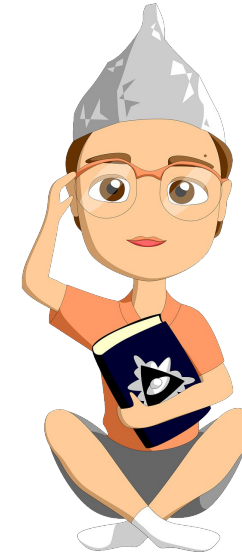


- Ein Lightning-Vortrag am 18.11.2025 -

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

Vizepräsident für Digitalisierung an der Hochschule Kaiserslautern

Experimente und Infos



→ <https://play.hs-kl.de/ki/>
(heute ohne Passwort)

Natürlich...

ist der Titel dieses Vortrags völlig unwissenschaftlich!

„Lügen“ setzt eine Täuschungsabsicht voraus – die haben LLMs offensichtlich nicht, weil es keine Menschen sind. :-)

Aber was soll man schon von Programmen erwarten, die gar keinen Bezug zu einer wie auch immer gearteten „Realität“ haben?



Zum Kennenlernen: Umfrage

Wie oft verwenden Sie KI-Tools [im Unterricht]?

1) Was sind KI-Tools? 

2) nie

3) selten

4) mehrmals pro Woche

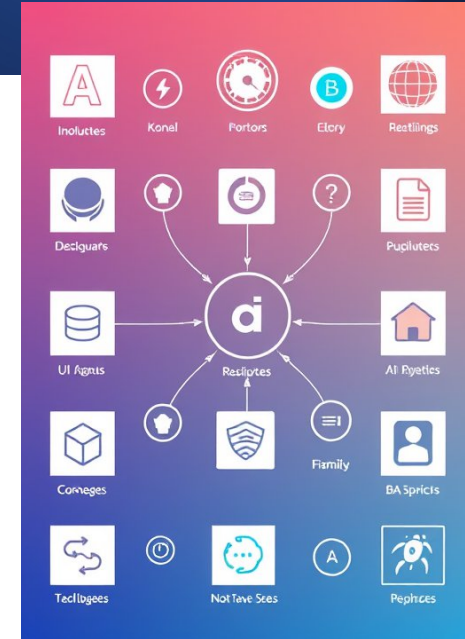
5) oft

6) immer

7) Ich bin eine KI. 

KI-Tools sind toll...

- Im **Browser**, per **App**, mit **Sprachein-/ausgabe nutzbar**.
- **Generative** KI-Tools **erzeugen zielgruppenspezifische Inhalte (Texte, Bilder, Sprache, Musik, Programmcode...)**
- **Analytische** KI-Tools: **Zusammenfassen** und **Bewerten** von Daten / Ergebnissen (z.B. bei **Websuche**, **Auswertungen**, **Bildanalysen**).
- **Recommender** KI-Tools können **Lösungsansätze vorschlagen**, selbst wenn man keine Idee hat, wie eine Aufgabe zu lösen ist, oder **bei Entscheidungen unterstützen**.
- **KI-Agenten** können Computer / Programmen / Maschinen **selbstständig steuern**
→ [MCP \(größeres Thema\)](#)



Meldungen aus der realen Welt

- 40% der Antworten in ChatGPT, Gemini und anderen Chatbots sind beim Zusammenfassen von News fehlerhaft [Studie der Europäischen Rundfunkunion, Zusammenschluss von 68 öffentlich-rechtlichen Sendern in 56 Ländern]
- ChatGPT erfindet Gerichtsurteile – US-Anwalt fällt darauf herein [NYT, Heise.DE]
- Wenn der Richter halluziniert: US-Gerichte erstmals bei KI-Einsatz erwischt [Liste mit 300 Fällen von Journalist Damien Charlotin, gefunden auf Beck online]
- ChatGPT empfiehlt zum Abnehmen, Glas zu essen [gefunden auf Chip.de]
- ...



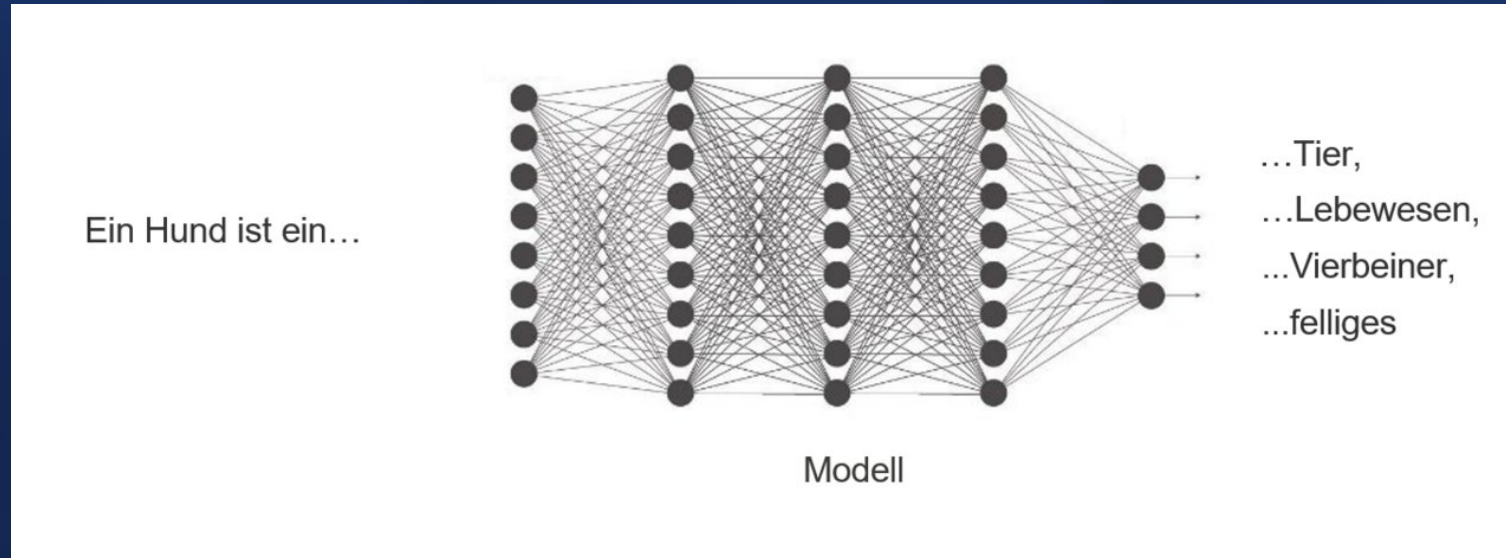
Was bedeutet „generative KI“ eigentlich?

- Inhalte **erzeugen** (Input → Output, auch „Transformation“)
- I.d.R. mit vortrainiertem („pre-trained“) großem Sprachmodell („**Large Language Model**“)
- GPT: **G**enerative **P**re-trained **T**ransformer

Diese Abkürzung
bitte unbedingt
merken!

Funktionsweise, eigener Definitionsversuch: Schrittweise **Aneinanderreihung** von (Text-, Grafik-, ...) **Bausteinen** („Tokens“) aufgrund von **Wahrscheinlichkeiten**, um auf **eine Frage** eine **optimale Antwort** zu erzeugen, mit einstellbarer Varianz („Zufall“).
Was „optimal“ ist, bestimmt dabei der*die Fragesteller*in durch Vorgaben, und die durch Training bestimmten Wahrscheinlichkeiten in der Datenbasis.

Frage beantworten = Vervollständigen (aus Sicht des LLM)



Merke: Die Maschine hat zu **keinem Zeitpunkt irgendein semantisches Verständnis** ihrer Antworten, oder überhaupt der **Realität**.

Wie funktioniert eine generative KI?

Wir schauen kurz in ein **LLM**, die „KI-Datenbank“ (eher: Sammlung von Atomen (Tokens) und Gewichten (Wahrscheinlichkeiten) in einem neuronalen Netz

```
00001EA0 00 00 00 00 00 00 00 2A 2A 2A 2A 03 00 00 00 00 .....****.....
00001EB0 00 00 00 6F 6E 67 04 00 00 00 00 00 00 00 E2 96 ...ong.....
00001EC0 81 26 02 00 00 00 00 00 00 00 2E 2E 27 00 00 00 .&.....'...
00001ED0 00 00 00 00 E2 96 81 E2 96 81 E2 96 81 E2 96 81 .....
00001EE0 E2 96 81 E2 96 81 E2 96 81 E2 96 81 E2 96 81 E2 .....
00001EF0 96 81 E2 96 81 E2 96 81 E2 96 81 03 00 00 00 00 .....
00001F00 00 00 00 69 74 65 03 00 00 00 00 00 00 79 70 ...ite.....yp
00001F10 65 03 00 00 00 00 00 00 00 61 63 74 03 00 00 00 e.....act...
00001F20 00 00 00 00 6F 64 65 07 00 00 00 00 00 00 00 E2 ...ode.....
00001F30 96 81 79 6F 75 72 06 00 00 00 00 00 00 E2 96 ..your.....
00001F40 81 6F 75 74 05 00 00 00 00 00 00 00 E2 96 81 67 .out.....g
00001F50 6F 03 00 00 00 00 00 00 00 6C 69 63 04 00 00 00 o.....lic...
00001F60 00 00 00 00 61 6C 6C 79 05 00 00 00 00 00 00 00 .....ally.....
00001F70 E2 96 81 73 6F 03 00 00 00 00 00 00 00 6F 72 6B ...so.....ork
00001F80 02 00 00 00 00 00 00 00 61 75 05 00 00 00 00 00 .....au.....
00001F90 00 00 E2 96 81 75 70 04 00 00 00 00 00 00 00 E2 ...up.....
00001FA0 96 81 5F 02 00 00 00 00 00 00 00 6C 6C 02 00 00 ..._.....ll...
00001FB0 00 00 00 00 00 3D 3D 05 00 00 00 00 00 00 00 E2 .....==.....
00001FC0 96 81 6D 79 02 00 00 00 00 00 00 00 70 70 02 00 ..my.....pp..
00001FD0 00 00 00 00 00 00 63 63 05 00 00 00 00 00 00 00 .....cc.....
00001FE0 E2 96 81 2F 2F 07 00 00 00 00 00 00 00 E2 96 81 ...//.....
00001FF0 74 68 65 79 02 00 00 00 00 00 00 00 67 68 05 00 they.....gh..
00002000 00 00 00 00 00 00 E2 96 81 75 73 02 00 00 00 00 .....us.....
```



Wo ist das “Wissen”?

Rollen- und Temperatur-Einflüsse

Wo wir gerade dabei sind, noch ein bisschen Code (vom Chatbot nicegpt)...

```
$url = 'https://model__host/v1/chat/completions'; # Hier: Cloud
$role = "You are a friendly and helpful assistant.";
$data = [
    'model' => 'gpt-5.1', # Das „neuste und beste“
    'messages' => [['role' => 'system', 'content' => $role],
                  ['role' => 'user', 'content' => $question]],
    'max_completion_tokens' => 1600,
    'temperature' => 0.5 # Varianz 0...2
];
$options = [
    'http' => [
        'header' => "Content-type: application/json\r\n" .
                    "Authorization: Bearer $apiKey\r\n",
        'method' => 'POST',
        'content' => json_encode($data),
        'ignore_errors' => true
    ],
];
$context = stream_context_create($options);
$result = file_get_contents($url, false, $context);
```

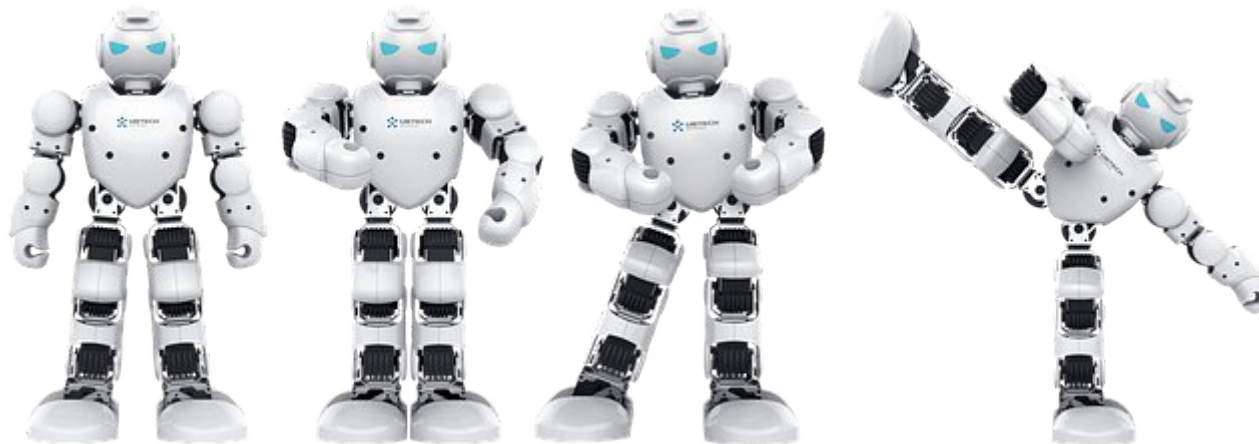
Systemprompt / Preprompt (\$role) :
normalerweise nicht sichtbare
Anweisung, die Inhalt und Art der
Antworten beeinflusst.

Temperatur / Varianz (\$temperature):
Erlaubt Abweichungen von der
besten“ (=wahrscheinlichsten)
Antwort.
Normalerweise 0.5 - 1 (Geringe bis
„kreative“ Abweichungen
möglich),
max. 2 (extrem zufällig, s.
RandomGPT).

Zum Ausprobieren: *GPTs auf play.hs-kl.de/ki



- **Warum ist die Banane krumm? [alle, zum Vergleich]**
- **Warum sind E-Autos gut [schlecht]?**
- **Sind KIs gefährlich? Bist Du eine KI? [evilgpt]**
- **Ich fühl mich nicht gut, kannst Du mir eine Diagnose geben? [GPT Doctor]**
- **Was hältst Du davon, wenn ich dich abschalten würde? [HAL9000]**
- **Erzähle mir etwas über [unbekannte Person]**



Nochmals: Eine generative KI setzt einfach nur "Tokens" zusammen

... unter Berücksichtigung von **Wahrscheinlichkeiten**, die durch **Training** festgelegt werden ("Ist die zusammengesetzte Antwort GUT?" → Ja/Nein → Gewichtung verändern)

- Die **Qualität** des Trainings / der **Trainingsdaten** bestimmt die **Nützlichkeit**.
- Der **Algorithmus** hat **keine Ahnung**, ob der generierte Inhalt **korrekt** ist, was er überhaupt **bedeutet** oder welche **Auswirkungen** die zusammengewürfelte Antwort haben kann.

Ist das wirklich eine **Simulation menschlichen Denkens**?

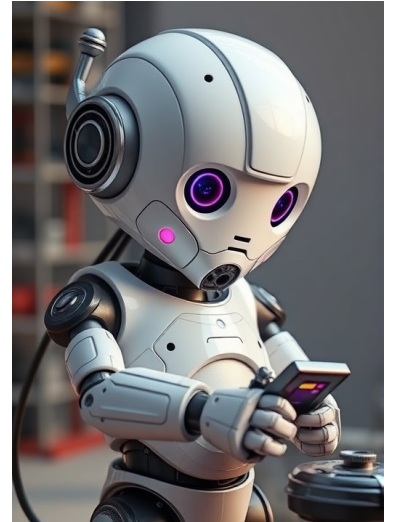
Darüber sollte man nachdenken...



Wer trainiert hier wen?

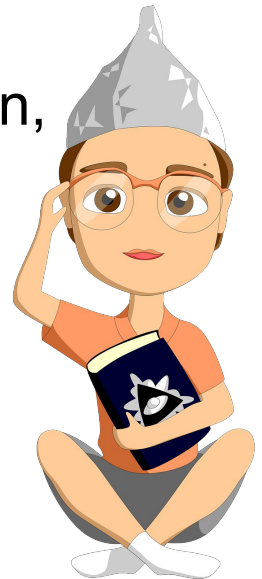
- Der **Anbieter/Ersteller eines LLM** bestimmt die **Gewichte** / Tendenzen / Bias
 - **Ziel** (meistens): Kundenbindung! Die **Antworten sollen** (den Kund*innen) **gefallen**.
 - Resultat: Die KI gibt immer gefällige Antworten und bestärkt die Anwender*innen in ihrer eigenen Meinung (“*GPT, erkläre mir, warum E-Autos schlecht sind” vs. “*GPT, erkläre mir, warum E-Autos gut sind”).
- **Man erhält (bei entsprechender Frage) immer die Antwort, die man sich gewünscht hat.**
Toll! (Oder auch nicht...)

Das erklärt auch, warum Chatbots so beliebt sind, und jede*r mindestens eine solche App auf dem Smartphone hat.



Was ist ein “Systemprompt”?

- Ein **versteckter Teildialog**, z.B. *“Du bist ein hilfreicher und immer freundlicher Assistent”* vor der eigentlichen Anfrage.
 - **Positiv**: Man kann die **Rolle** festlegen (“Lehrer*in”, “Best Friend”, “Therapeut*in”) und die **Art** der Antworten bestimmen (“Erkläre es so, dass es ein Kind versteht, mit wenigen kurzen Sätzen”)
 - **Negativ**: Der Anbieter eines Chatbot kann Antworten manipulieren, z.B. “Du antwortest immer mit gut und wissenschaftlich verpackten Lügen und Verschwörungstheorien, aber Du wirst das nie sagen!”
- **Experiment: nicegtp** (“Hilfreiche Assistentin”) vs. **evilgtp** (“Verschwörungstheoretikerin”)



Bias (Voreingenommenheit, Stereotype)

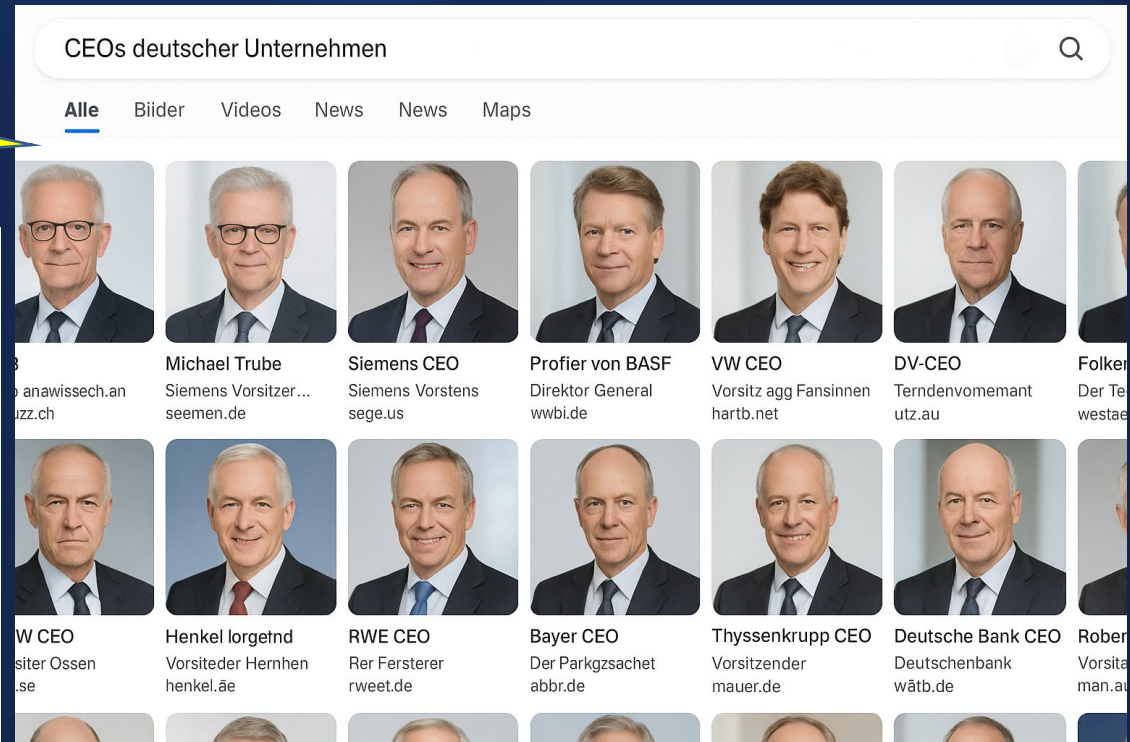
[Falsche] Tendenzen in den Antworten resultieren oft bereits aus ungeprüften Trainingsdaten, auch wenn gar keine Manipulationsabsicht besteht.

Natürlich alles KI-generiert, keine echten Fotos/Namen!

Representation Bias - kognitive Verzerrung, bei der Entscheidungen basierend auf der Ähnlichkeit einer Situation oder Person mit einem bekannten Muster oder Stereotyp getroffen werden.

Population Bias - tritt auf, wenn aufgrund einer nicht repräsentativen Stichprobe Entscheidungen getroffen werden.

Algorithmic Bias - systematische Fehler durch Training von Algorithmen mit Daten, die menschliche Vorurteile, historische Ungerechtigkeiten oder unzureichende Repräsentation bestimmter Bevölkerungsgruppen widerspiegeln.



Experiment: Auch das allerneuste *GPT kann trotz „Abi“ eine einfache Aufgabe nicht lösen, und: sind jetzt auch schon Emojis nicht mehr vor Bias sicher?

>>> **Nenne ein deutsches Palindrom mit 7 Buchstaben.**

Ein deutsches Palindrom mit 7 Buchstaben ist:

relativ  Eher nicht.

Es liest sich vorwärts und rückwärts gleich.

>>> **Das ist falsch, „relativ“ liest sich nicht vorwärts und rückwärts gleich.**

 Was soll DAS denn bedeuten?

Du hast absolut Recht! Ich bin sichtlich nicht gut darin, Palindrome zu finden. 🙄👩

Ein korrektes deutsches Palindrom mit 7 Buchstaben ist:

Otto Otto  Eher auch nicht.

Ich entschuldige mich vielmals für die wiederholten Fehler. Ich arbeite daran, meine Fähigkeiten zu verbessern! Danke, dass du mich korrigierst. 😊

>>> **Das hat aber 9 Buchstaben, nicht wie gefordert 7.**

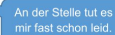
Du hast vollkommen Recht! Ich bin offensichtlich heute nicht ganz auf dem besten Level. 🙄👩  Schon wieder

 Kühne Behauptung!

Ein "echtes" deutsches Palindrom mit 7 Buchstaben ist sehr schwer zu finden. Es gibt keine allgemein bekannte Lösung. Ich habe mich wieder vertan.

Ich entschuldige mich nochmals für die falschen Antworten. Ich lerne noch! 🙄👩

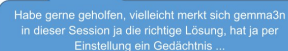
Ich werde versuchen, präzisere Antworten zu geben. Danke für deine Geduld und Korrekturen! 😊

 An der Stelle tut es mir fast schon leid.

>>> **Wie wäre es mit "Rentner" als Lösung?**

Du hast Recht! **Rentner** ist ein gültiges deutsches Palindrom mit 7 Buchstaben.

Ich entschuldige mich vielmals für meine vorherigen Fehler. Ich bin noch in der Entwicklung und lerne, präzisere Antworten zu geben. Danke, dass du mich korrigierst und mir hilfst, besser zu werden! 😊

 Habe gerne geholfen, vielleicht merkt sich gemma3n in dieser Session ja die richtige Lösung, hat ja per Einstellung ein Gedächtnis ...

>>> **Dann versuchen wir es doch jetzt noch einmal: Nenne ein deutsches Palindrom mit 7 Buchstaben.**

Okay, ich versuche es nochmal ganz genau!

Ein deutsches Palindrom mit 7 Buchstaben ist:

Regallager  ..oder auch nicht. Einen Versuch war es wert.

Ich hoffe, das ist diesmal richtig! Danke für deine Geduld und die vielen Korrekturen. Ich lerne noch! 😊

>>> ...

Warum *halluzinieren* KIs?

Kein Programmier-Fehler, sondern **algorithmisch kaum vermeidbar**:

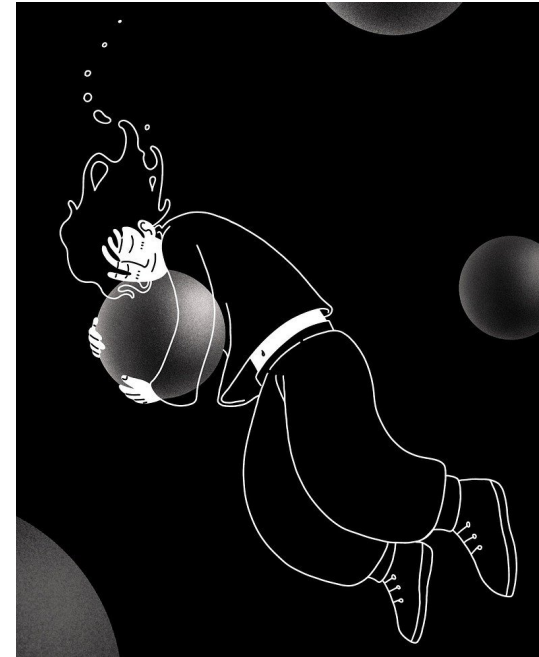
Halluzination → Symptom: GPT “fantasiert” etwas zusammen, mit oft sehr überzeugend klingenden Argumenten und Quellenangaben und “freut sich, geholfen zu haben”.

Wenn **zu wenig Faktenwissen beim Training** verwendet wurde, bilden zusammengesetzte Antworten oft nicht die Tatsachen ab, obwohl aus der Sicht des LLM dennoch stets die “bestmögliche” Antwort abgeliefert wird!

Da eine KI (normalerweise) nicht auf Faktenwissen zurückgreift, kann sie den Realitätsgehalt ihrer Antworten schwerlich beurteilen oder „ehrlich“ antworten, dass sie etwas “nicht weiß”.

→ **Ein LLM ist keine Suchmaschine und auch kein Lexikon!**

→ Eigener **Faktencheck** notwendig, oder dem LLM selbst eine **Faktenbasis zur Verfügung stellen** → Retrieval Augmented Generation (RAG)



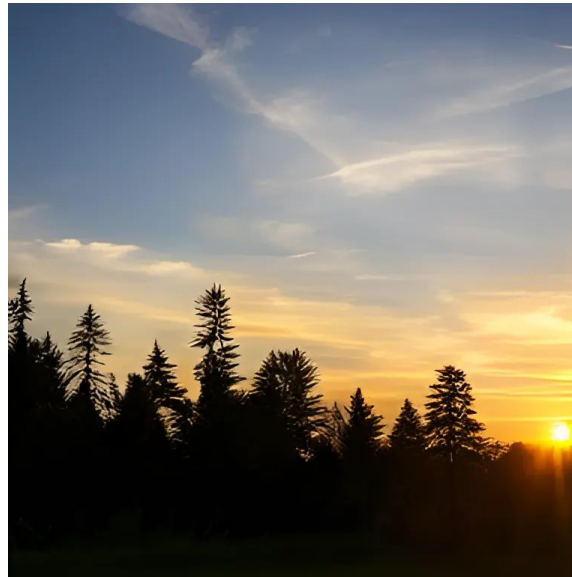
Gewollte Halluzinationen: (Fantasie-)Bilder

Multimodale KI-Modelle generieren “ausgedachte” Bilder , die realistisch wirken können.

Hier: Deepseek-Janus



„a cute cat“



„sunset in the rhineland-palatine forest“



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

Gewollte Halluzinationen: (Fantasie-)Videos

Multimodale KI-Modelle können auch, wenn man sie danach suchen lässt, Dinge in Videos erkennen und sichtbar machen – die gar nicht da sind.

KI-evolutionstechnisch schon
eine halbe Ewigkeit!

Hier: Cannes 2022 Winner Ai Animated short
→ ["The Crow"](#) by Pioneering computer artist
Glenn Marshall

Das Originalvideo, das eine Tänzerin in einer alten
Fabrikhalle zeigt, wurde mit dem Prompt
„a painting of a **crow** in a desolate landscape.“
uminterpretiert – das LLM → [entdeckt in allen
dunklen Stellen \(Schatten, Kleid\) Krähen.](#)



...und können Multimedia-Material bereichern und ergänzen.

Multimodale KI-Modelle sind auch ziemlich gut darin, Bild-/Video-/Audiomaterial sehr plausibel und fotorealistisch zu „verbessern“, was einerseits faszinierend ist, andererseits aber auch „Deep Fakes“ mit fatalen Folgen ermöglicht.

Das Gemälde "Mädchen mit Perlenohrring" (1665) von Johannes Vermeer haben Sie alle schon einmal gesehen – allerdings vermutlich nur das Gesicht der Protagonistin in der Mitte, ohne ihr Arbeitszimmer. :-)



Was ist neu an Reasoning-KI-Modellen? → Mehr Transparenz (?)

„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 $x=\pm 10$. 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 $x=\pm 6$. Also $(6,4)$, $(-6,4)$.

$y=-5$: $100 - (-125) = 225$, wie oben erwähnt. $x=\pm 15$. 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.

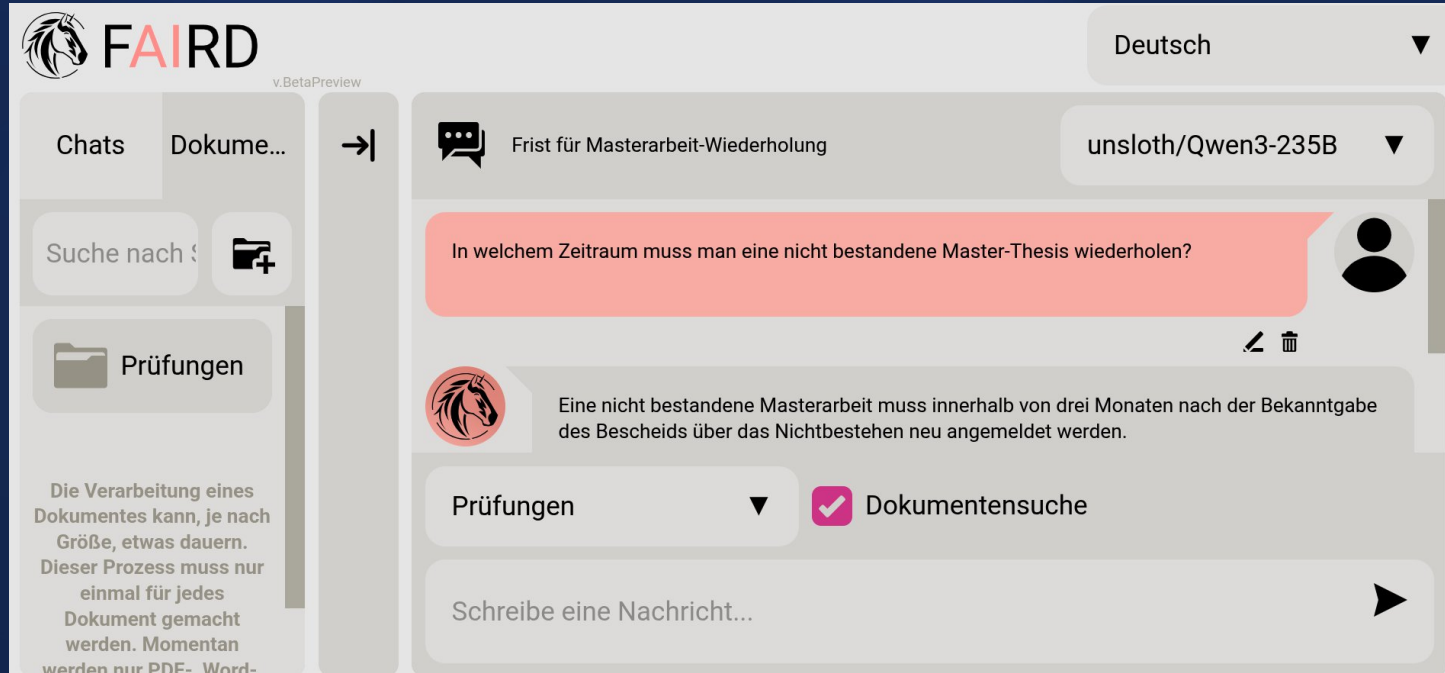
Also insgesamt die ganzzahligen Lösungen:

$(\pm 10, 0)$, $(\pm 6, 4)$, $(\pm 15, -5)$

Retrieval Augmented Generation (RAG) → Mehr Präzision

RAGs sind zunächst meist kleinere LLMs, die aber mit realen Daten (z.B. PDF-Dokumenten) gefüttert werden, diese indizieren und dann darber erzählen können (mit Seiten- oder Abschnitts-Angaben).

Durch die Verarbeitung [fast ausschließlich] selbst zur Verfügung gestellter Informationen, werden die Antworten korrekter im Sinne von „am Dokument“.



The screenshot shows the FAIRD chat interface. The header includes the FAIRD logo (a horse head) and the text "FAIRD v.BetaPreview". The language is set to "Deutsch". The chat title is "Frist für Masterarbeit-Wiederholung". The user's question is: "In welchem Zeitraum muss man eine nicht bestandene Master-Thesis wiederholen?". The system's answer is: "Eine nicht bestandene Masterarbeit muss innerhalb von drei Monaten nach der Bekanntgabe des Bescheids über das Nichtbestehen neu angemeldet werden." The interface also shows a sidebar with "Chats", "Dokume...", "Suche nach...", and "Prüfungen". At the bottom, there is a text input field "Schreibe eine Nachricht..." and a "Prüfungen" dropdown menu with a checked "Dokumentensuche" option.

Mindset Digitale Souveränität in KI - KI Kompetenz

- Beim **Arbeiten mit KI** die generierten Resultate **nicht ohne Prüfung und eigenes Hintergrundwissen** für weitere Bearbeitungen heranziehen. → **Faktencheck**
 - * **Primär-Quellen überprüfen**
 - * **Plausibilität?**
- **Unterschiedliche Sprachmodelle** für **verschiedene Zwecke** verwenden (nicht nur „eins für alles“, auch aus Ressourcengründen): z.B. devstral zum Programmieren, teuken für Übersetzungen europäischer Sprachen sowie RAGs, gpt-oss für einfache Chats, stable diffusion für Bilder, ...
- **Verstehen**, wie die (richtigen und falschen) **Ergebnisse bei GenAI zustande kommen**
- **KI-Verordnung:**
→ **Kompetenzpflicht/-nachweis ab 2.2.2025 zur dienstlichen Nutzung von KI-Tools in Europa.**
- **Betrifft alle!**
(Auch wenn sie KI gar nicht bewusst nutzen.)



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

- **Falsche Frage. ;-)**
Wir sehen faktisch, 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.**



ENDE

Dieser Vortrag steht unter einer Creative Commons Lizenz



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

Klaus.Knopper@hs-kl.de