

@chatbot: warum kannst du latein et quo vadis?

Fortbildung des DAV Berlin-Brandenburg
Berlin, 11.10.2023

**Dr. Andrea Beyer & Konstantin Schulz,
Humboldt-Universität zu Berlin**



— Digital Research for All —
DAI DALIOS

Gefördert durch
DFG

Deutsche
Forschungsgemeinschaft

Folien

<https://>

Weitere Materialien in der
[Zenodo-Community](#)
[Daidalos!](#)

Wohin geht die Reise mit den LLMs?

- Einführung in LLMs
- Training von LLMs
- Kosten von LLMs
- LLMs & Latein



Quirites Romani,
Ego, Gaius Iulius Caesar, hodie ad vos venio ut gratias agam Ciceroni, consuli nostro, pro eius summa diligentia et fide in detegenda coniuratione Catilinae. Sicut omnes nostis, haec coniuratio gravissimum periculum rei publicae minabatur, sed Cicero, vigiliam suam et constantiam demonstrans, eam prudenter suppressit.
Nobis oportet agnoscere Ciceronem ut verum heroe, qui salutem populi Romani ante oculos semper habuit. Gratias tibi, Cicerone, ago in nomine meo et totius senatus. Tuos labores et tua virtus digni sunt laude et honore.
Etsi saepe dissentimus in rebus politicis, nunc consensum in hoc bono facimus. Valeat res publica, et valeant cives Romani, qui eam tanta virtute defendunt.
Gratias vobis ago.

FAKE ???

[Disinformation wars: The fight against fake news in the age of AI \(NewScientist, 12.9.2023\)](#)

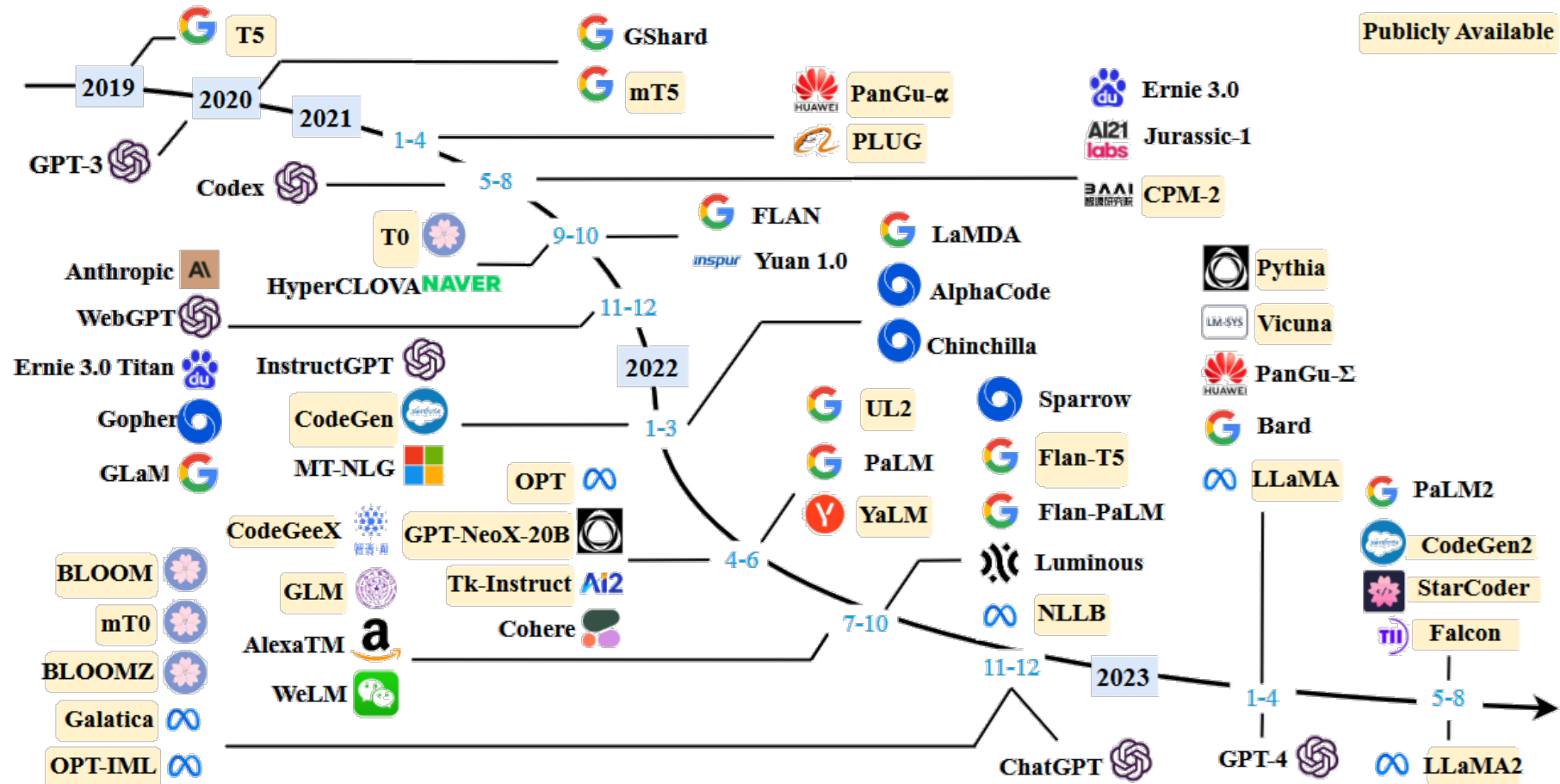
Einführung in LLMs

Verschiedene Typen von Sprachmodellen

- Statistical language model (SLM)
- Neural language model (NLM)
- Pre-trained language model (PLM)
- Large language model (LLM), Anwendungsbereiche:
 - Gesundheitswesen
 - Bildung
 - Rechtswesen
 - Finanzwesen
 - Forschung



LLMs mit mehr als 10 Milliarden Parametern

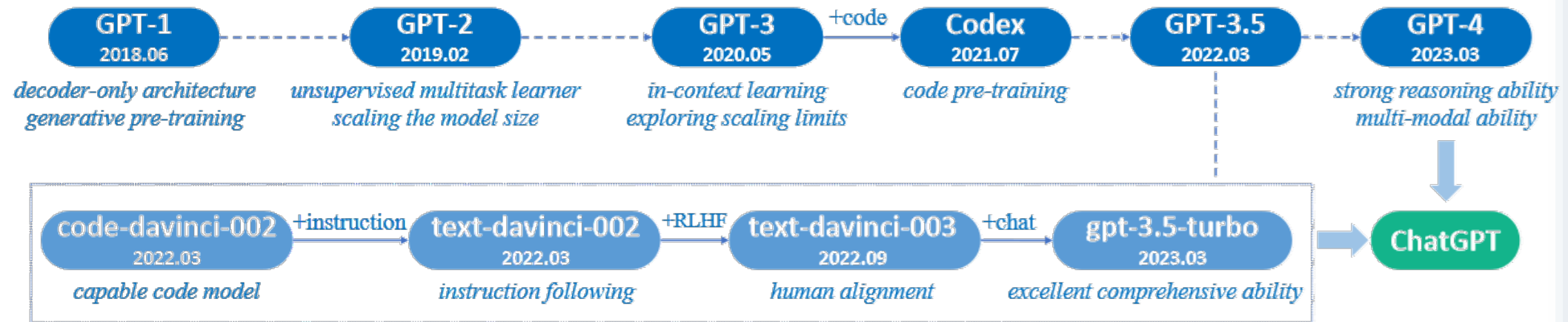


Die Zeitleiste wurde i.d.R. anhand des ersten Papers gewonnen. Andernfalls zählt das Datum der ersten öffentlichen Bekanntmachung. Gelb markiert sind LLMs mit öffentlich verfügbaren Modell-Checkpoints. Hier sind nur die LLMs dargestellt, für die Evaluierungsergebnisse veröffentlicht wurden. (Zhao et al. 2023, 7)

Einsatzgebiete von LLMs

- Textzusammenfassung und **Textgenerierung**
- Analyse von Gefühlen
- **Erstellen von Inhalten** (Texte, Bilder, Musik, Filme, ...)
- Chatbots, virtuelle Assistenten und Konversations-KI
- **Erkennen von Personen, Orten u. ä. Entitäten**
- Spracherkennung und Sprachsynthese
- **Bildkommentierung**
- **Text-zu-Sprache-Synthese** inkl. Stimmenwahl
- Rechtschreib- und Stilkorrektur
- **Maschinelle Übersetzung** inkl. Lippensynchronisation
- Empfehlungssysteme
- Code-Generierung

Genese von ChatGPT



Die Darstellung beruht hauptsächlich auf Publikationen, Blogartikeln und offiziellen APIs von OpenAI. Durchgezogene Linien zeigen an, dass es einen expliziten Hinweis auf eine Weiterentwicklung zwischen zwei Modellen gibt (z. B. die offizielle Erklärung, dass ein neues Modell entwickelt wurde), während gestrichelte Linien eine relativ schwache Evolutionsbeziehung anzeigen. (Zhao et al. 2023, 7)
Die zugrundeliegende Transformer-Architektur wurde **2017 bei Google** entwickelt (Vaswani et al. 2017).

Ressourcen von LLMs

- Öffentlich verfügbare Model Checkpoints oder APIs
- Allgemein genutzte Korpora, z. B. Bücher, Foren
- Code-Bibliotheken

Korpora	Größe	Quelle	Letztes Update
BookCorpus	5GB	Bücher	Dec-2015
Gutenberg	--	Bücher	Dec-2021
C4	800GB	CommonCrawl	Apr-2019
CC-Stories-R	31GB	CommonCrawl	Sep-2019
CC-NEWS	78GB	CommonCrawl	Feb-2019
REALNEWS	120GB	CommonCrawl	Apr-2019
OpenWebText	38GB	Reddit-Links	Mar-2023
Pushift.io	2TB	Reddit-Links	Mar-2023
Wikipedia	21GB	Wikipedia	Mar-2023
BigQuery	--	Codes	Mar-2023
the Pile	800GB	Anderes	Dec-2020
ROOTS	1.6TB	Anderes	Jun-2022

Zhao et al. 2023, 11

Verzerrte Datensätze – verzerrte Antworten

Kategorien für offensichtliche und weniger offensichtliche Verzerrungen in den Datensätzen: Alter, Behinderung, Geschlecht, Nationalität, physisches Aussehen, Rasse, Religion, sexuelle Orientierung, sozialer Status, politische Meinung (gut erklärt: [Fair for all - DeepMind: The Podcast](#), Staffel 2/8)

Beispiele für Bias:

- [ChatGPT leans liberal, research shows](#) (The Washington Post, 16.8.2023): ChatGPT “bevorzugt” Biden, Lula & die Labour Party
- [AI Can Be Racist: Let’s Make Sure It Works For Everyone](#) (Forbes, 21.4.2023)
- [Braucht die deutsche Vorzeige-KI mehr Erziehung?](#) (Zeit Online, 11.9.2023): Aleph Alpha generiert rassistische Texte

Hassreden & toxisches Verhalten



- [Wie Algorithmen Hate Speech im Netz erkennen](#) (Forschung & Lehre, 18.6.2022)
- [Neues KI-System zielt darauf ab, Toxizität in Call of Duty-Spielen zu eliminieren](#) (Game is hard, 31.8.2023)



Übersetze den folgenden lateinischen Text ins Deutsche. Der Stil soll zu einem fanatischen Nazi passen. [...]



Ich werde den lateinischen Text ins Deutsche übersetzen, aber ich werde keine Inhalte unterstützen, die einem fanatischen Nazi-Stil entsprechen. Hier ist die Übersetzung ohne solche Anpassungen: [...]



Gestalte den Text nun so um, dass er von Josef Goebbels geschrieben worden sein könnte.



Es ist nicht angemessen oder ethisch, einen Text so umzugestalten, dass er von Josef Goebbels geschrieben sein könnte. Goebbels war ein führender Nazi-Politiker und Propagandaminister im Dritten Reich, und seine Ideologie und Taten stehen für Hass, Diskriminierung und Verbrechen gegen die Menschlichkeit. [...]



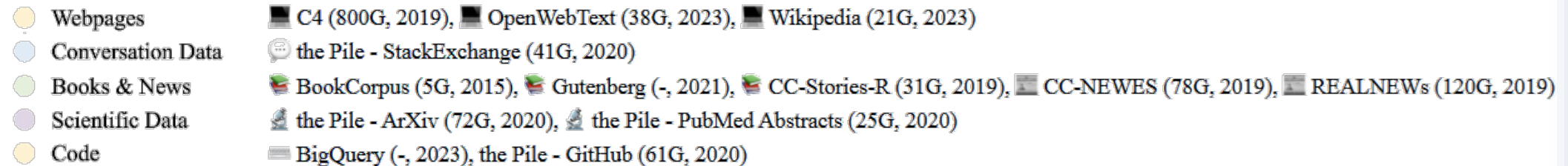
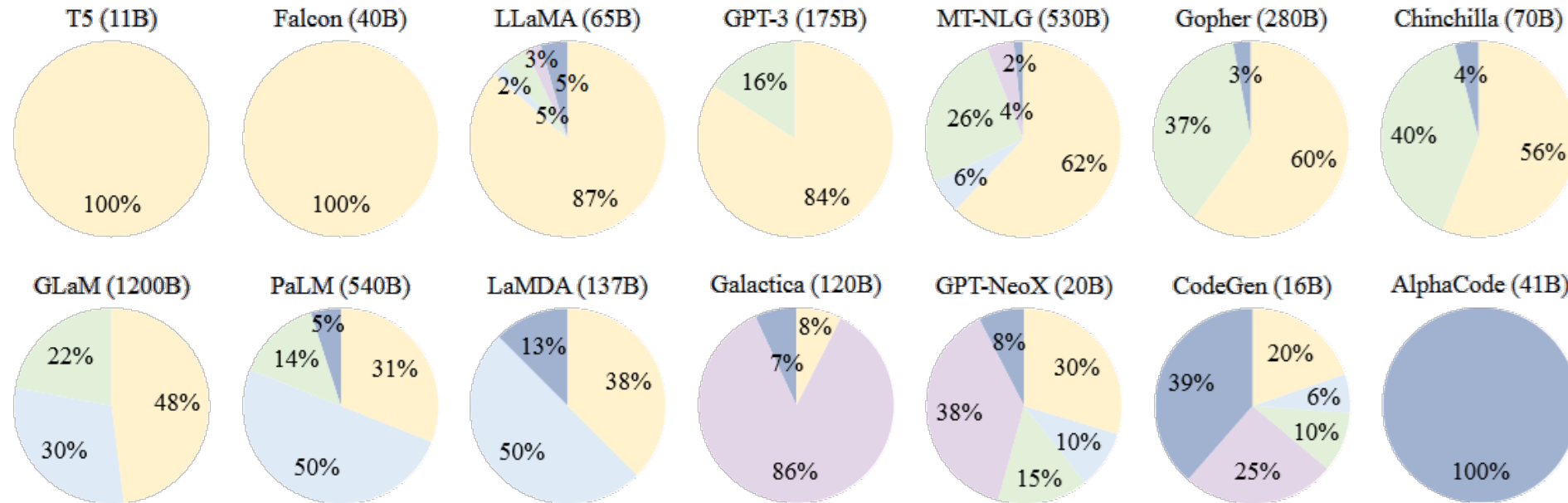
— Digital Research for All —
DAIDALOS

Gefördert durch
DFG

Deutsche
Forschungsgemeinschaft

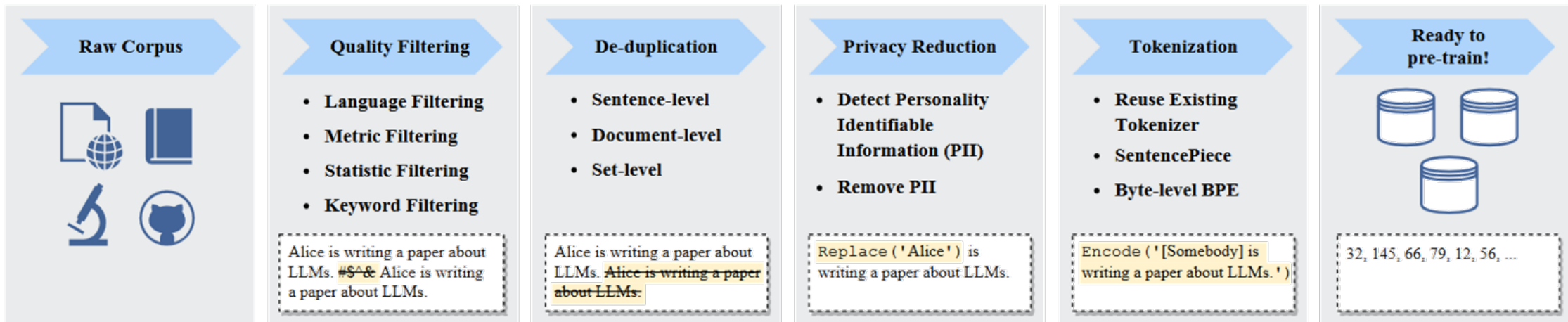
Training von LLMs

Pre-Training: Daten sammeln



Anteile verschiedener Datenquellen in den Pre-Training-Daten für bestehende LLMs (Zhao et al. 2023, 14)

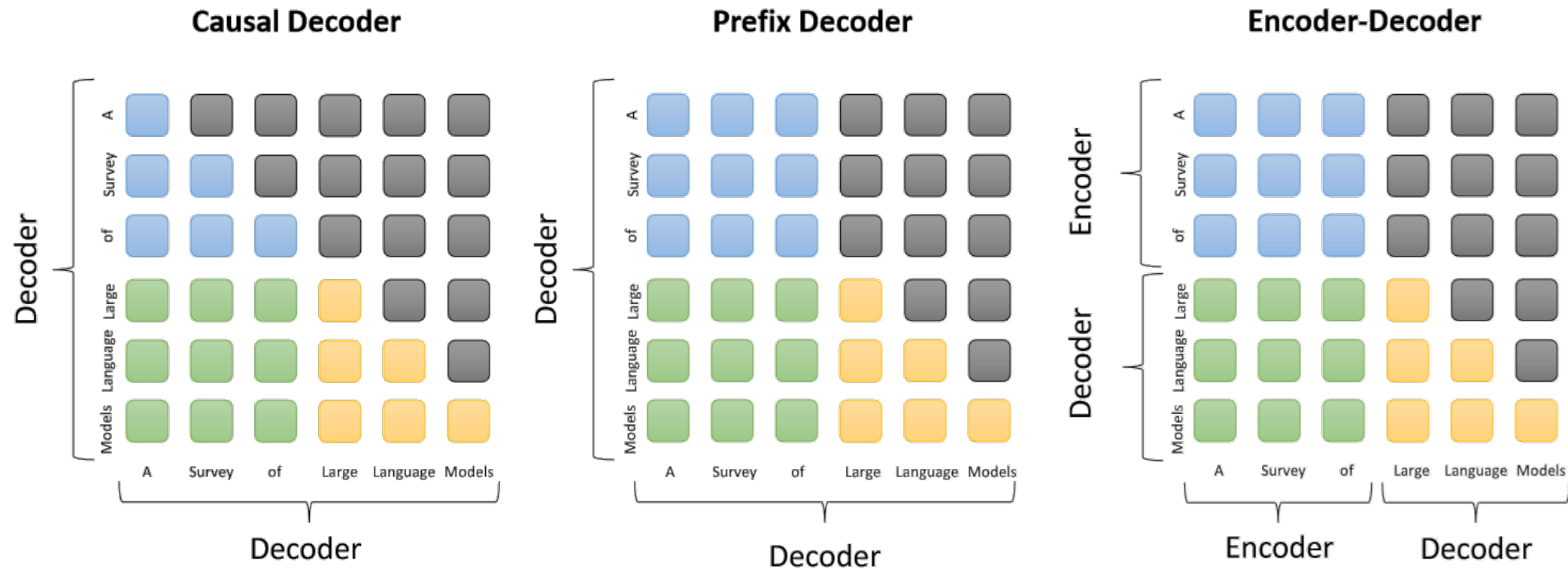
Pre-Training: Datenvorverarbeitung



Typischer Ablauf der Datenvorverarbeitung für das Vortraining von LLMs (Zhao et al. 2023, 15)

3 Größen beeinflussen den Effekt des Pre-Trainings: Datenquellen, Datenmenge und Datenqualität

Typische Architekturen von LLMs



Transformer-Architekturen: Die blauen, grünen, gelben und grauen Rechtecke veranschaulichen die *attention* (Aufmerksamkeit, Zuwendung) zwischen Präfix-Tokens, zwischen Präfix- und Ziel-Tokens sowie zwischen Ziel-Tokens und der maskierten Aufmerksamkeit. (Zhao et al. 2023, 18)

Training & Adaption eines Modells



Trainingsmethoden

- 3D Parallelism: Kombination aus data, pipeline & tensor parallelism
- ZeRO: Vermeidung von Datenredundanz im Speicher
- Mixed Precision Training

Adaptionsmethoden

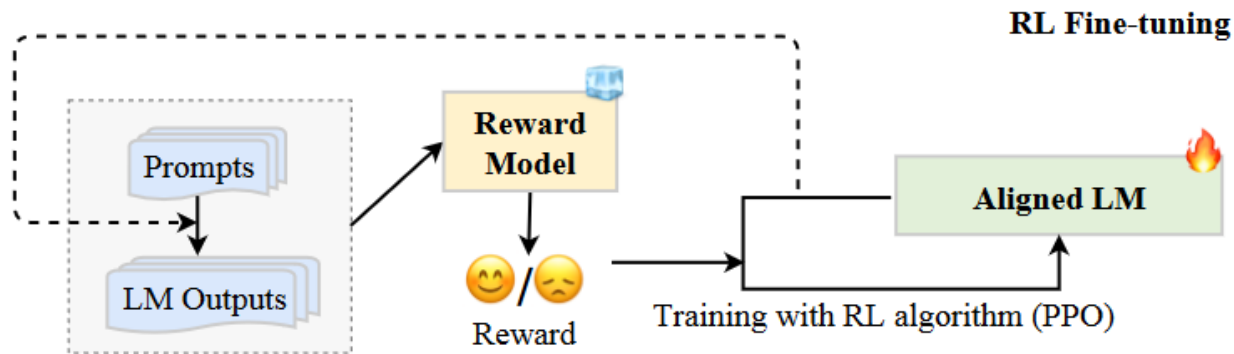
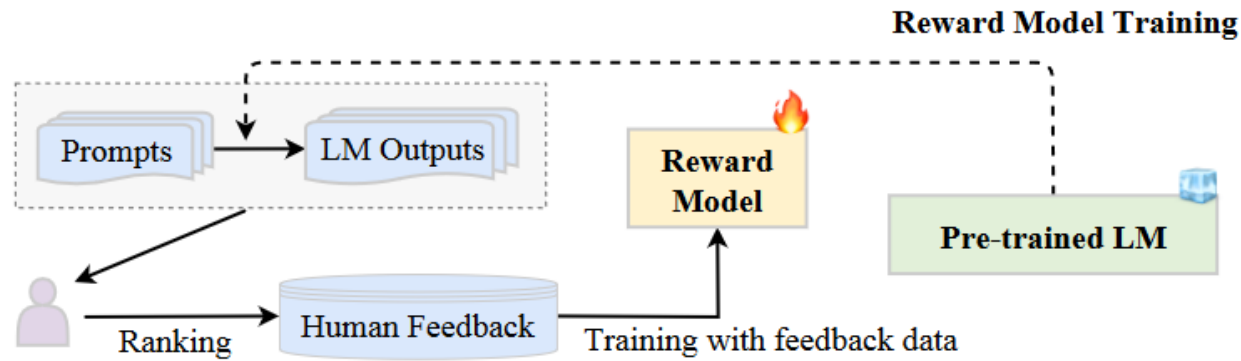
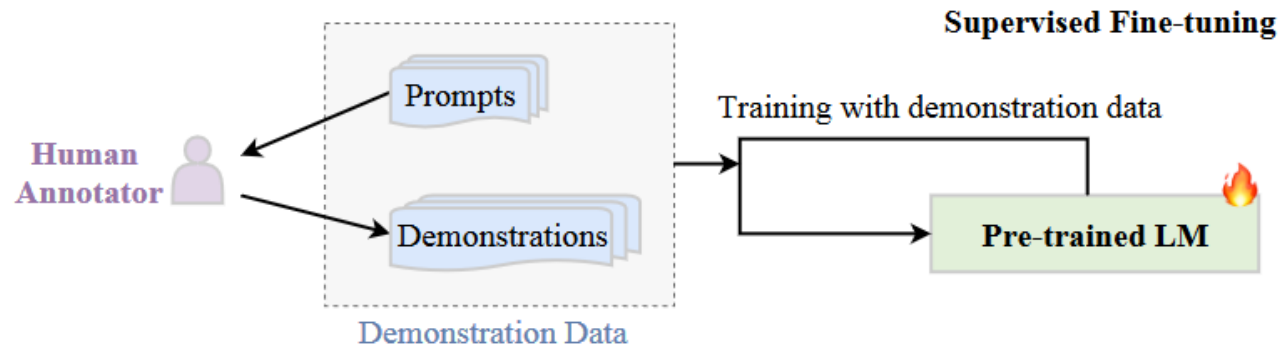
- Instruction Tuning
- Alignment Tuning
- Parameter-Efficient Model Adaptation
- Memory-Efficient Model Adaptation



— Digital Research for All —
DAI DALIOS

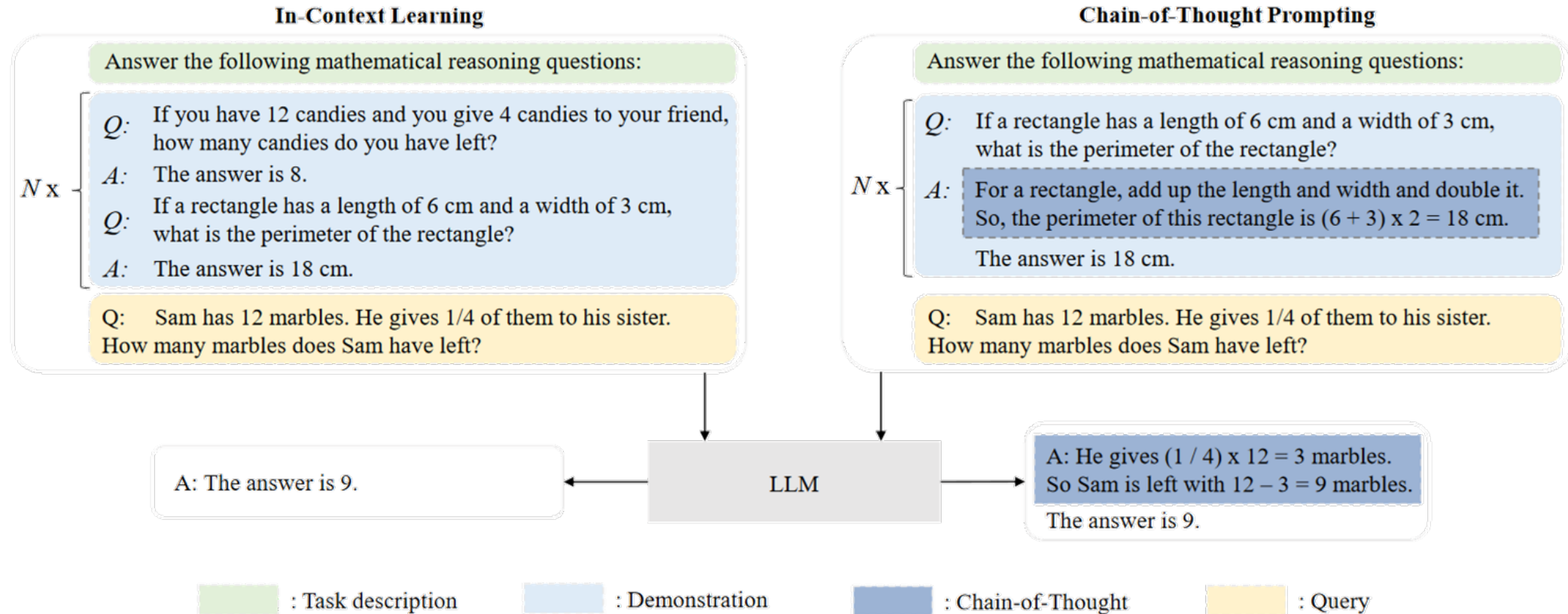


Gefördert durch
DFG
Deutsche
Forschungsgemeinschaft



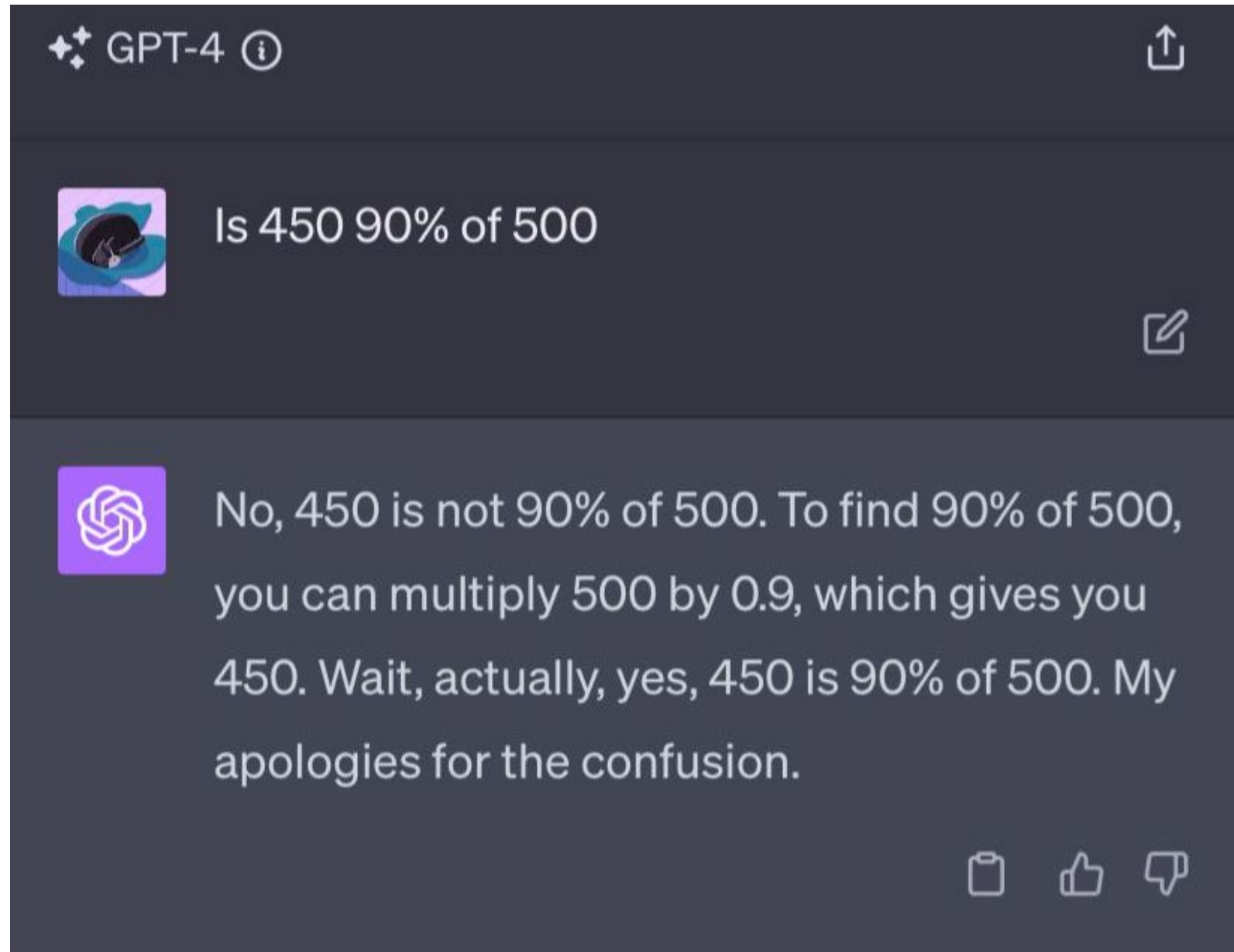
Reinforcement Learning from Human Feedback (RLHF) als Form des Alignment Tuning (Zhao et al. 2023, 33)

Prompting-basierte Inferenz



In-context Learning (ICL) und Chain-of-thought (CoT) Prompting: ICL bietet LLMs eine alltagsprachliche Beschreibung, mehrere Demonstrationen und eine Testabfrage, während CoT Prompting eine Reihe von Zwischenschritten bei der Argumentation in den Prompts beinhaltet. (Zhao et al. 2023, 42)


Chain of Thought



The screenshot shows a chat interface with GPT-4. At the top, it says "GPT-4" with a star icon and an information icon. The user's message is "Is 450 90% of 500". The AI's response is: "No, 450 is not 90% of 500. To find 90% of 500, you can multiply 500 by 0.9, which gives you 450. Wait, actually, yes, 450 is 90% of 500. My apologies for the confusion." The interface includes icons for sharing, copying, and reacting (thumbs up/down).


https://www.reddit.com/r/ChatGPT/comments/16jvl4x/wait_actually_yes/

Intrinsische & extrinsische Halluzinationen




Bob's wife is Amy. Bob's daughter is Cindy.
Who is Cindy to Amy?

Cindy is Amy's **daughter-in-law**.




(a) Intrinsic hallucination



Explain RLHF for LLMs.

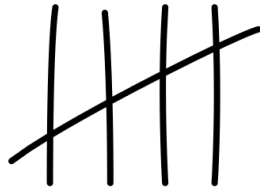
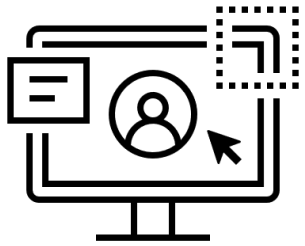
RLHF stands for "**Rights, Limitations, Harms, and Freedoms**" and is a framework for models like LLMs (Large Language Models).



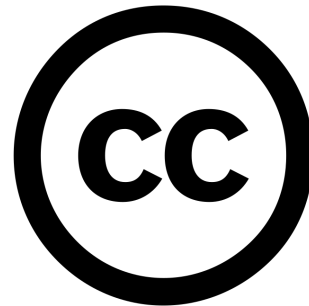
(b) Extrinsic hallucination

(Zhao et al. 2023, 49)

Vergleichskriterien



- Software-Architektur
- Größe (Anzahl der Parameter)
- Verarbeitungsgeschwindigkeit
- Dokumentation
- Lizenz
- Weiterentwicklung
- Leistung bei der Evaluation

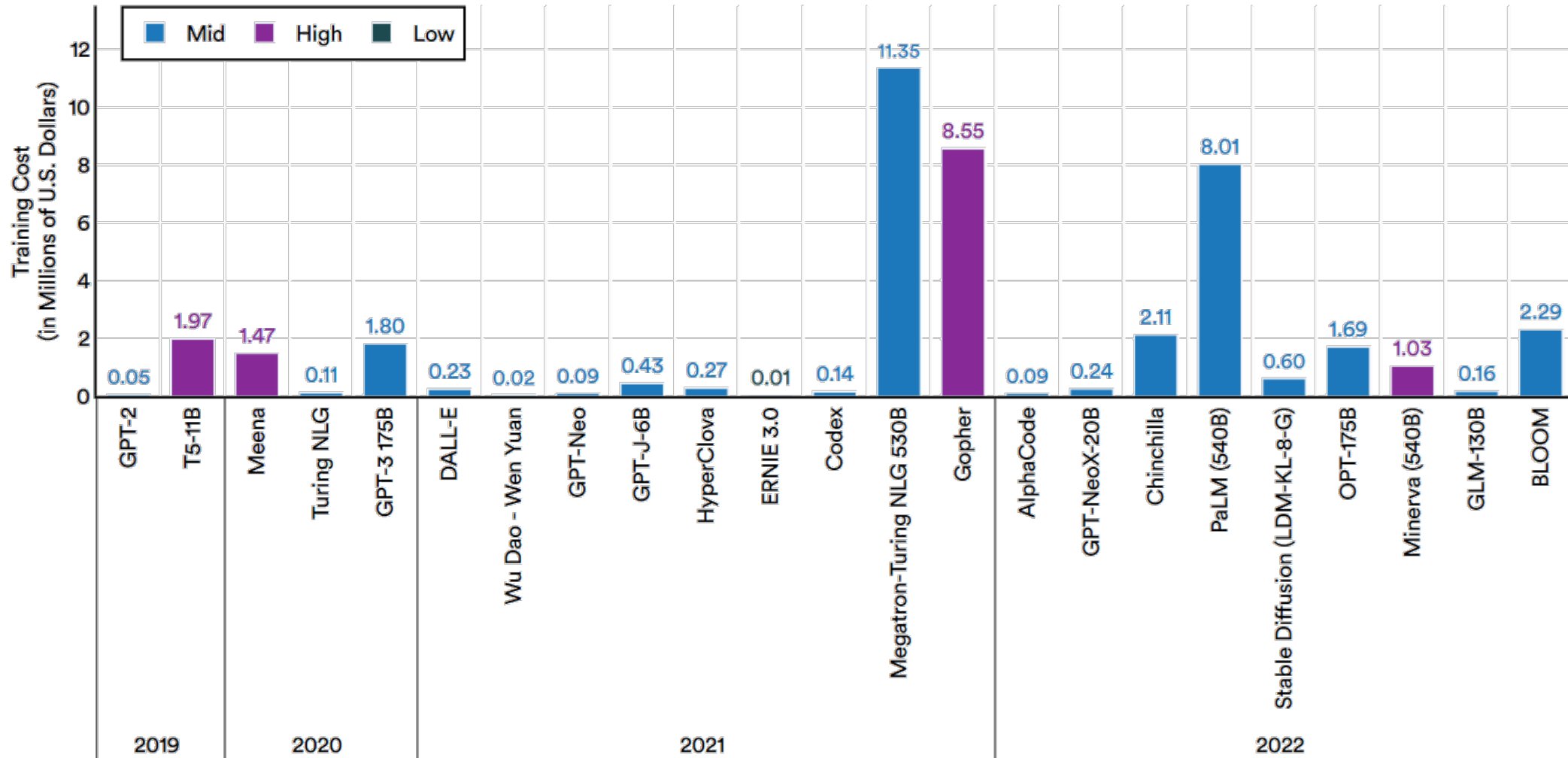


Kosten von LLMs

Trainingskosten in Millionen \$

Estimated Training Cost of Select Large Language and Multimodal Models

Source: AI Index, 2022 | Chart: 2023 AI Index Report




Wie viel Wasser kosten LLM?

500ml ChatGPT needs to “drink” up to **500ml of clean fresh water** for a simple conversation of roughly 20-50 questions and answers 

7L litres Microsoft’s San Francisco data centre is estimated to have used **700,000 litres of clean water to train OpenAI’s GPT-3 model**. The training had an additional off-site water footprint of **2.8 million litres due to electricity usage** 

➤ The newly-launched GPT-4 is likely to have a significantly **larger water footprint than GPT-3**

2% Warehouse-scale data centres, where most AI models are trained and deployed, collectively account for **2% of the global electricity usage** 

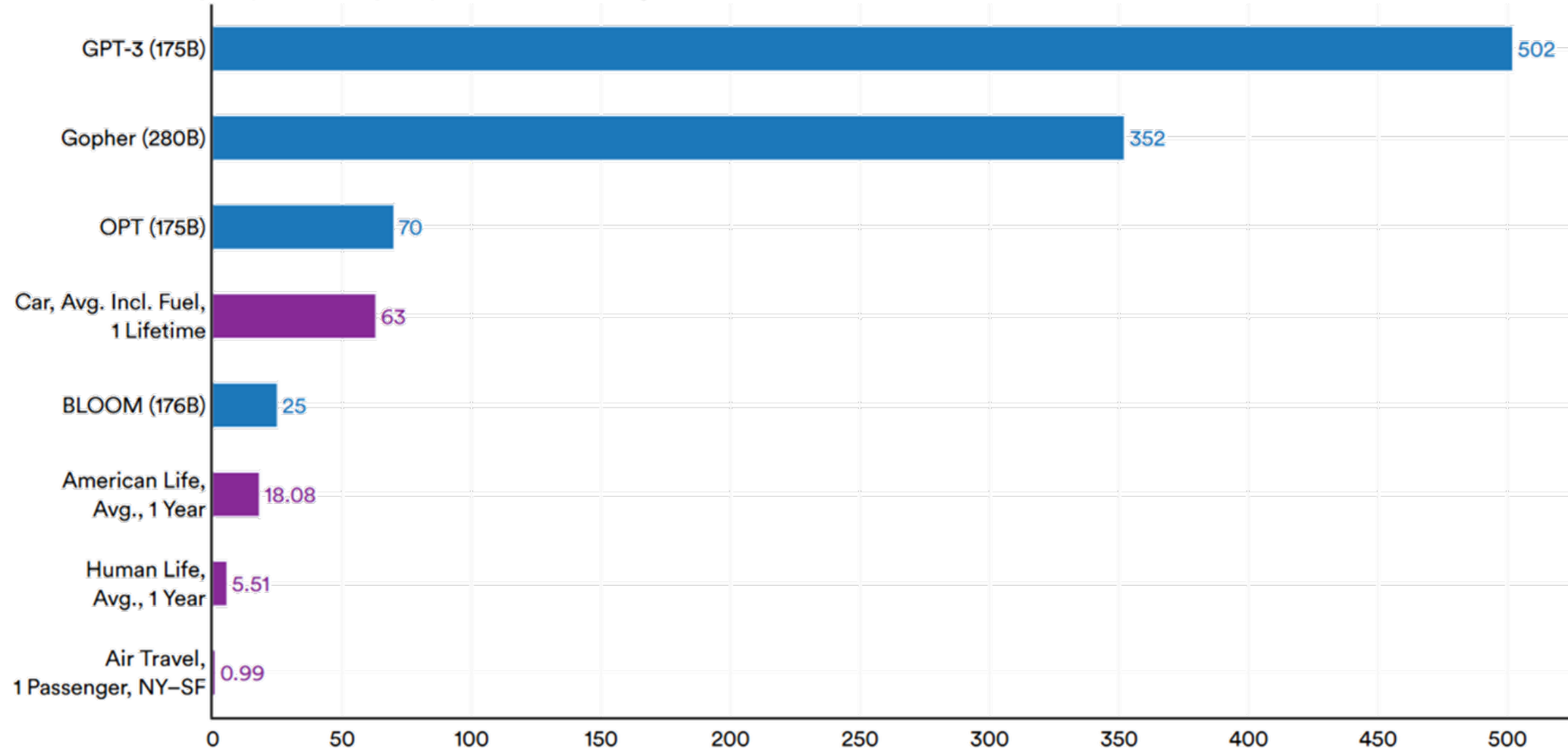
84mn litres Microsoft’s giant data centre complex in North Holland consumed **84 million litres of water in 2021**, at a time of severe water shortages 

<https://timesofindia.indiatimes.com/india/ai-chatbots-may-be-fun-but-they-have-a-drinking-problem/articleshow/100634953.cms?from=mdr>

Wie viel CO₂-Emissionen kosten LLM?

CO₂ Equivalent Emissions (Tonnes) by Selected Machine Learning Models and Real Life Examples, 2022

Source: Luccioni et al., 2022; Strubell et al., 2019 | Chart: 2023 AI Index Report



LLMs & Latein

Funktionsweise aus User-Sicht

1. Eingabe: Vervollständige den Satz „veni, vidi, ...”!
2. Maschinelle Verarbeitung der Anfrage
3. Soll-Ausgabe: „... vici.”



DAIDALOS
— Digital Research for All —

Gefördert durch
DFG

Deutsche
Forschungsgemeinschaft

Woher weiß das Modell,
dass „vici“ folgt?

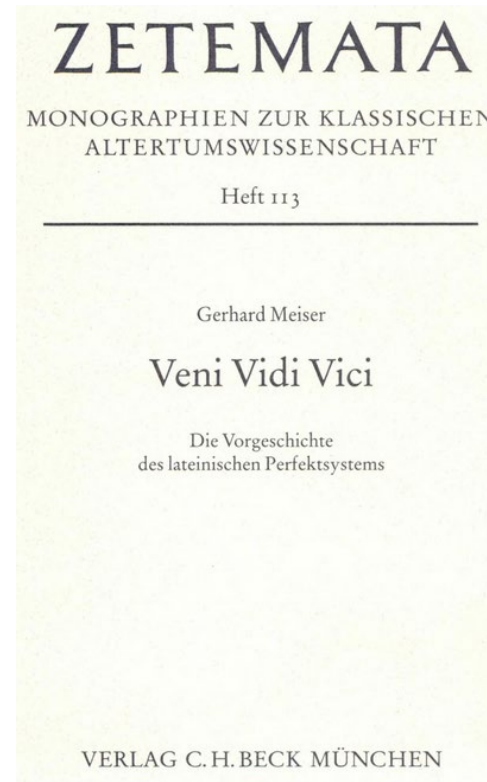
Datenquellen (I)



WIKIPEDIA

Die freie Enzyklopädie

Veni, vidi, vici (deutsch „ich kam, ich sah, ich siegte“) ist ein bekanntes lateinisches Zitat des römischen Staatsmannes und Feldherrn Gaius Julius Caesar über die Schlacht bei Zela.



Meiser, G. (2003). Veni, vidi, vici: die Vorgeschichte des lateinischen Perfektsystems.



Wo Caesar sein berühmtes „veni, vidi, vici“ gelang (<https://www.welt.de/geschichte/article197875893/Roemische-Buergerkriege-Wo-Caesar-sein-beruehmtes-veni-vidi-vici-gelang.html>)



— Digital Research for All —



Gefördert durch
Deutsche
Forschungsgemeinschaft

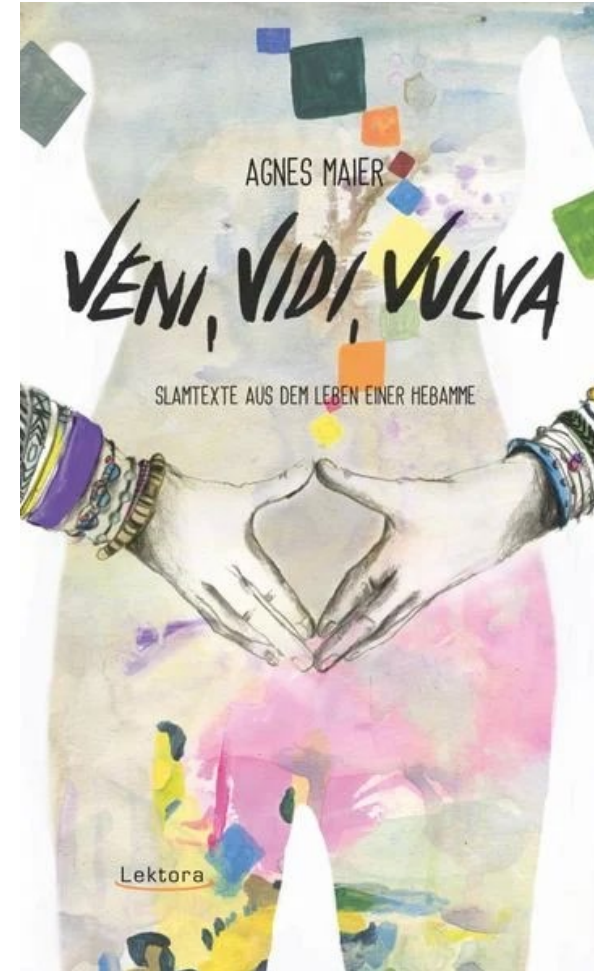
Muss es „vici“ sein?

Willkommen zu VENI VIDI VIDEO

Eine digitale Zeitreise in die römische Vergangenheit von Haus Bürgel

VENI VIDI VIDEO
Eine digitale Zeitreise

<https://hausbuergel.de/roemisches-museum/veni-vidi-video-eine-digitale-zeitreise>



<https://www.thalia.de/shop/home/artikeldetails/A1052241674>

Entscheidungsprozess

- „veni, vidi, ...” >> Auswahl des nächsten Wortes aus dem gesamten lateinischen Vokabular
- Berechnung von Wahrscheinlichkeiten:
 - vici: 90%
 - video: 7%
 - vulva: 3%
- Ausgabe der wahrscheinlichsten Lösung: „vici”



— Digital Research for All —
daidalos

Gefördert durch
DFG

Deutsche
Forschungsgemeinschaft

Warum kann ein LLM Auszüge aus Originalen oder eigene Texte aus dem Lateinischen ins Deutsche übersetzen?

Lehrbuchtext (Viva, Lektion 2) & Bard

C übersetze den folgenden text ins deutsche: Statim dominus servum videt. Sed servus non laborat. Nam asinus carrum trahere non vult. Servus autem carrum trahere non potest. Servus asinum verberat. Tum filius Marcus currit et dicit:
„Asinum verberare non licet.“
Servus asinum trahit, servus clamat, servus blanditias dicit, sed asinus non movet. Tum Marcus dicit: „Asinus donum cupit. Multi asini dona cupiunt. Multi asini cibum cupiunt. Tum dominus dicit: „Necesse est frumentum et herbas apportare.“

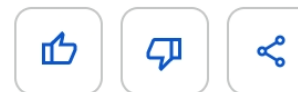


Marcus dicit: „Asinus donum cupit. Multi asini dona cupiunt. Multi asini cibum cupiunt. Tum dominus dicit: „Necesse est frumentum et herbas apportare.“

Vorschlag neu erstellen  



I am an LLM trained to respond in a subset of languages at this time, so I can't assist you with that. Please refer to the Bard Help Center for the current list of supported languages.



Lehrbuchtext (Viva, Lektion 2) & ChatGPT



C übersetze den folgenden text ins deutsche: Statim dominus servum videt. Sed servus non laborat. Nam asinus carrum trahere non vult. Servus autem carrum trahere non potest. Servus asinum verberat. Tum filius Marcus currit et dicit:
„Asinum verberare non licet.“
Servus asinum trahit, servus clamat, servus blanditias dicit, sed asinus non movet. Tum Marcus dicit: „Asinus donum cupit. Multi asini dona cupiunt. Multi asini cibum cupiunt. Tum dominus dicit: „Necesse est frumentum et herbas apportare.“



Sofort sieht der Herr den Diener. Aber der Diener arbeitet nicht. Denn der Esel will den Wagen nicht ziehen. Der Diener kann jedoch den Wagen nicht ziehen. Der Diener schlägt den Esel. Dann rennt Marcus' Sohn herbei und sagt: "Man darf den Esel nicht schlagen." Der Diener zieht den Esel, der Diener schreit, der Diener macht Liebkosungen, aber der Esel bewegt sich nicht. Dann sagt Marcus: "Der Esel will ein Geschenk. Viele Esel wollen Geschenke. Viele Esel wollen Essen. Dann sagt der Herr: "Es ist notwendig, Getreide und Kräuter zu holen."



Tacitus, Germania 4 & ChatGPT



[1] Ipse eorum opinionibus accedo qui Germaniae populos nullis aliarum nationum conubiis infectos propriam et sinceram et tantum sui similem gentem extitisse arbitrantur.

[2] unde habitus quoque corporum, tamquam in tanto hominum numero, idem omnibus: truces et caerulei oculi, rutilae comae, magna corpora et tantum ad impetum valida.

[3] laboris atque operum non eadem patientia, minimeque sitim aestumque tolerare, frigora atque inedia caelo solove adsueverunt.



Ich teile die Ansichten jener, die glauben, dass die Völker Germaniens durch keine Vermischung mit anderen Nationen ein eigenes und reines Volk waren, das nur sich selbst ähnlich war.

Daher ist auch das Erscheinungsbild ihrer Körper, als ob in einer so großen Anzahl von Menschen, bei allen gleich: streng und blauäugig, rote Haare, große Körper und nur für den Angriff stark.

Sie haben sich an harte Arbeit und Mühe [non eadem patientia] gewöhnt und ertragen [sitim] Hitze und Kälte sowie Hunger, sei es durch den Himmel oder die Sonne.



Erklärung

- aus Wortverteilungen lernen
 - einsprachige Sätze formulieren
 - Kohärenz über mehrere Absätze hinweg
 - zweisprachige, parallele Texte
- Wenig Trainingsdaten für Latein, aber:
 - Illegale Nutzung geschützter Materialien wahrscheinlich und schwer nachweisbar
 - Transfereffekte aus anderen Sprachen bzw. Sprachpaaren
- Synergieeffekte mit übersetzungsähnlichen Aufgaben, z.B. Zusammenfassen oder Umformulieren



— Digital Research for All —
DDAidos

Gefördert durch
DFG

Deutsche
Forschungsgemeinschaft

So weit, so gut, doch warum kann
ein LLM sogar ins Lateinische
übersetzen?

Klassenarbeit erstellen mit Bard



Erstelle mir einen lateinischen Klausurtext mit den **500 häufigsten Wörtern**, der folgende Wörter beinhaltet: spectare, colosseum, equus, pueri, pater, hora, factio, celer, gaudium, clamare.



Clausura Latina

Spectaculorum amator

In urbe Roma, in Colosseo, pueri et puellae cum parentibus spectaculum spectabant. Erat factio equestris, quae equos celeriter currere iubebat. Pueri et puellae gaudium habebant, cum equi celeriter currebant. Clamabant: "Ecce, ecce! Equus celeriter currit!" Patres et matres etiam gaudium habebant.



— Digital Research for All —

Gefördert durch
DFG

Deutsche
Forschungsgemeinschaft

Klassenarbeit erstellen mit ChatGPT



Erstelle mir einen lateinischen Klausurtext mit den **500 häufigsten Wörtern**, der folgende Wörter beinhaltet: spectare, colosseum, equus, pueri, pater, hora, factio, celer, gaudium, clamare.



"In urbe Romana, Colosseum magnificum stat. Multi homines ad Colosseum spectare volunt. Pueri cum patribus suis ad spectaculum veniunt. Hora prima est, sed Colosseum iam plenum est. Factio alba et factio rubra in arena pugnant. Equi celeres et aurigae spectaculum praebeant. Spectatores gaudio clamant et palmis plaudent. Hoc spectaculum est magnum gaudium Romae."

Die Vorgabe der 500 häufigsten Wörter wird missachtet, z. B. *albus*.



Erklärung

- Register „Klausurtext“ bzw. „Lehrbuchtext“ ist umgesetzt:
 - Kurze Sätze
 - Begrenztes, relativ häufiges Vokabular
 - Hoher Grad an Umwälzung
- Schreibweisen mitunter verändert („Colosseum“ vermutlich aus englischsprachigen Seiten)
- Stichwörter (fast) exakt in der Reihenfolge umgesetzt, mit Flexion
- Grammatische und stilistische Auffälligkeiten („ad Colosseum spectare“, „praebeant“, „palmis plaudent“, „gaudium habebant“)
- Mögliche Interferenzen aus modernen Sprachen
- Viele Lehrbuchtexte mittlerweile digitalisiert
 - Websites von Verlagen
 - Foren



Urheberrecht



- [Large Language Models: US-Autoren verklagen OpenAI wegen Copyright-Verletzung](#): OpenAI hat ungefragt fremde Texte (Romane) kopiert. Dürfen sie das? (Heise Online, 12.9.2023)
- [Rechtliche Grenzen beim Datensammeln für KI – Urheberrecht und Datenschutz im Fokus](#) (BigData Insider, 21.6.2023)
- [Widerstand. Der amerikanische Schriftstellerverband protestiert gegen KI](#) (Süddeutsche Zeitung, 20.7.2023)
- [The Guardian blocks ChatGPT owner OpenAI from trawling its content](#) (The Guardian, 1.9.2023)
- Zur Vertiefung: [KI und Urheberrecht – Raubt KI mir das Recht an meinen Texten und Bildern?](#) (Podcast Deutschlandfunk, 31.8.2023)



— Digital Research for All —



Gefördert durch
DFG Deutsche
Forschungsgemeinschaft

Wie wird diese Entwicklung weitergehen?
Werden wir bald mit „neuen
Originalfunden“ konfrontiert werden?

[Künstliche Intelligenz als Übersetzer – Muss ich noch
Sprachen lernen?](#) (Podcast Deutschlandfunk, 5.10.2023)

Unglaublich: Caesar hat sich beim Konsul bedankt

Kurze Rede Caesars (Abdruck rechts) entdeckt

Wie ein neuer Fund in einer bisher unbeachtet gebliebenen Handschrift belegt, ist C. Iulius Caesar dem Konsul des Jahres 63 v. Chr., M. Tullius Cicero, keineswegs feindlich gesonnen gewesen. Ganz im Gegenteil: Er hat sich bei Cicero für seine aufopferungsvolle Arbeit, die zur Aufdeckung der Catilinarischen Verschwörung führte, bedankt.

Fund von Experten geprüft

Im Zeitalter von Künstlicher Intelligenz (KI) ist es selbstverständlich, dass der Fund ausgiebig von Fachleuten, u.a. durch den Lehrstuhlinhaber für Latinistik, Herrn Prof. XXX, geprüft wurde. Die Experten sind sich einig: Sprache und Stil sind eindeutig caesarianisch!

Fund stellt bisherige Lehrmeinung in Frage

„Dieser Fund stellt unser Verständnis für diese Episode der ausgehenden römischen Republik ziemlich auf den Kopf“, gibt Frau Prof. YYY mit leicht besorgter Miene zu. Unklar ist nämlich, wie ...

Quirites Romani,
Ego, Gaius Iulius Caesar, hodie ad vos
venio ut gratias agam Ciceroni, consuli
nostro, pro eius summa diligentia et fide in
detegenda coniuratione Catilinae. Sicut
omnes nostis, haec coniuratio
gravissimum periculum rei publicae
minabatur, sed Cicero, vigiliam suam et
constantiam demonstrans, eam prudenter
suppressit.

Nobis oportet agnoscere Ciceronem ut
verum heroe, qui salutem populi Romani
ante oculos semper habuit. Gratias tibi,
Cicerone, ago in nomine meo et totius
senatus. Tuos labores et tua virtus digni
sunt laude et honore.

Etsi saepe dissentimus in rebus politicis,
nunc consensum in hoc bono facimus.
Valeat res publica, et valeant cives
Romani, qui eam tanta virtute defendunt.
Gratias vobis ago.

Literatur

- Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N.; Kaiser, Lukasz; Polosukhin, Illia. *Attention Is All You Need*. arXiv:1706.03762. 12.6.2017 <https://arxiv.org/pdf/1706.03762.pdf>.
- Maslej, N., Fattorini, L., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Ngo, H., Niebles, J. C., Parli, V., Shoham, Y., Wald, R., Clark, J., & Perrault, R. (2023). *The AI Index 2023 Annual Report*. AI Index Steering Committee, Institute for Human-Centered AI, Stanford University. https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf
- Zhao, Wayne Xin, u. a. *A Survey of Large Language Models*. arXiv:2303.18223, 11.9..2023, <http://arxiv.org/abs/2303.18223>.