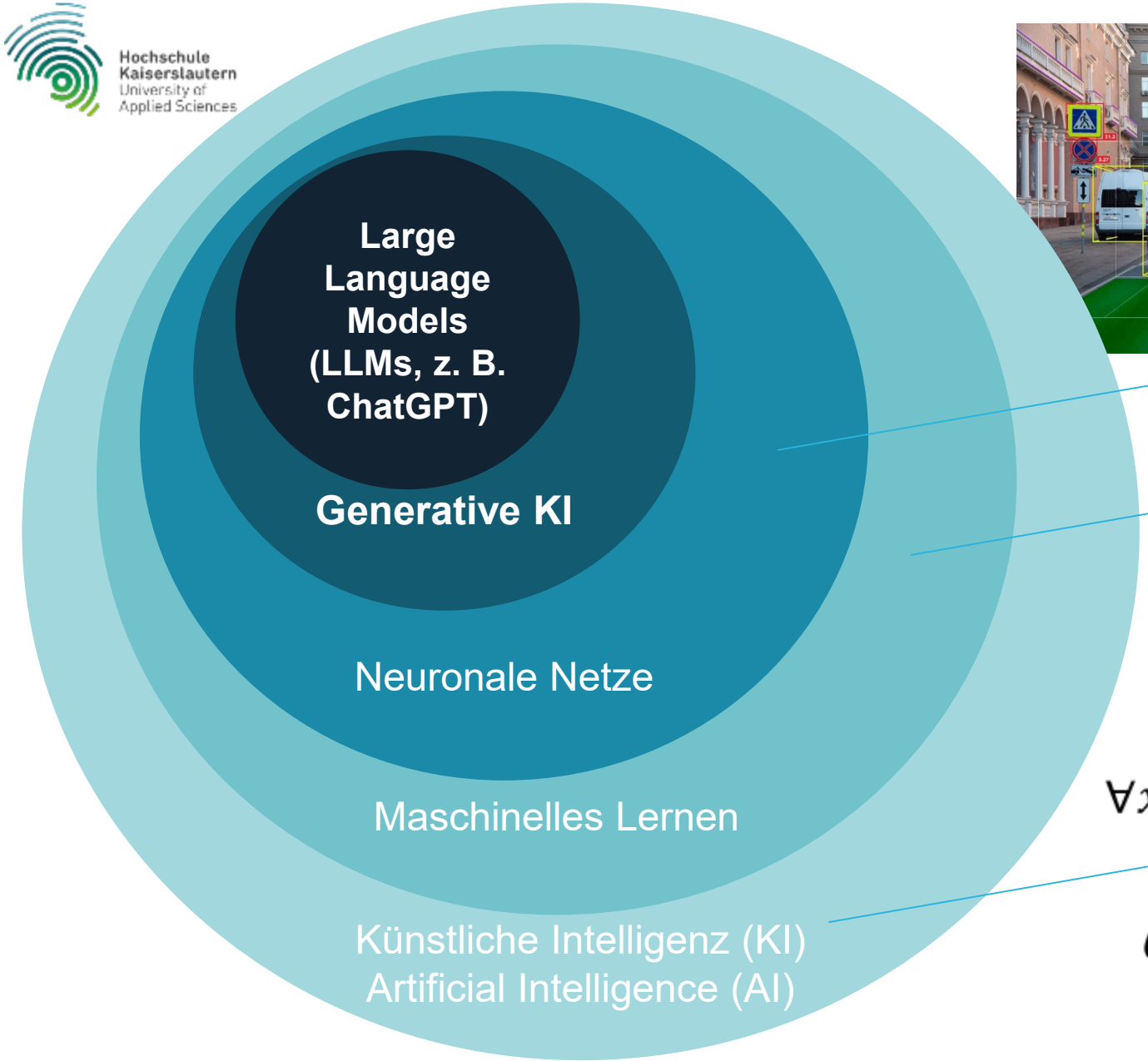
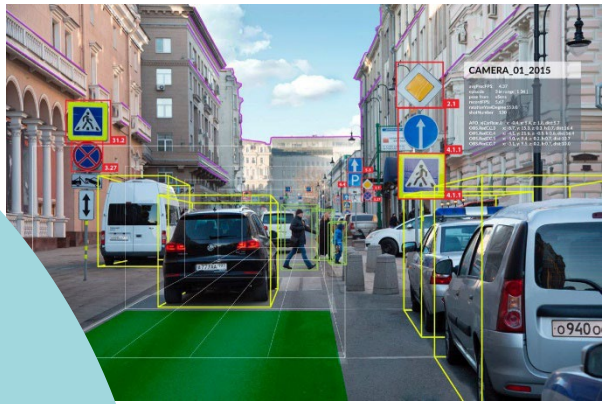
Hochschule Kaiserslautern
University of Applied Sciences

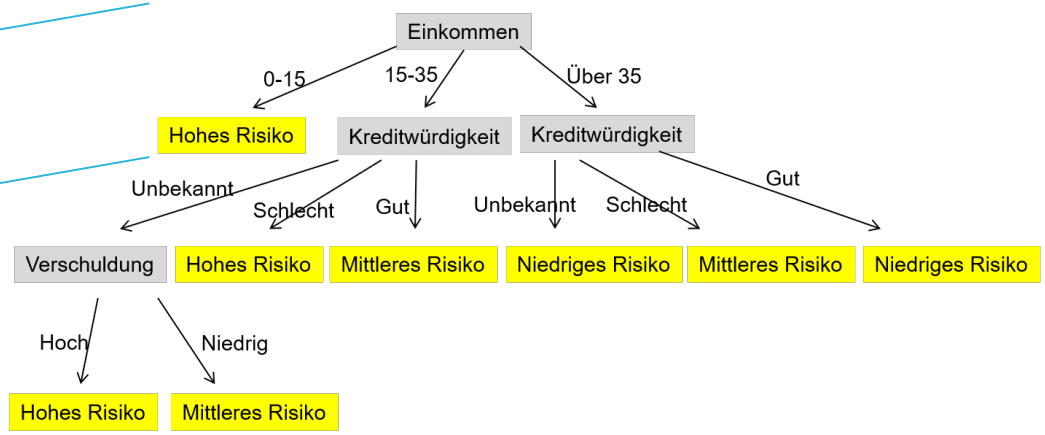
Wie intelligent ist Künstliche Intelligenz?

Symposium der Wirtschaftsinformatik,
Hochschule Kaiserslautern
23. Mai 2024

Prof. Dr. Eugen Staab
eugen.staab@hs-kl.de







$$\forall x\left[\forall y\left(Cat(x) \wedge Mouse(y)\right)\right] \Rightarrow Catch(x, y)$$

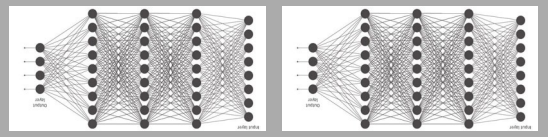
$$Cat(Tom);\ Mouse(Jerry)$$

Large Language Models (LLMs)

PROMPT

Erklär mir kurz, wie ein Elektromotor funktioniert.

Ein Elektromotor ist ein Gerät,



LLM (z. B. ChatGPT)

Trainiert mit ca. 1 Billion Wörter ≙ 10.000 Menschenleben

AUSGABE

Ein Elektromotor ist ein Gerät, …

Hochschule Kaiserslautern
University of Applied Sciences

Verständnis

Intelligenz

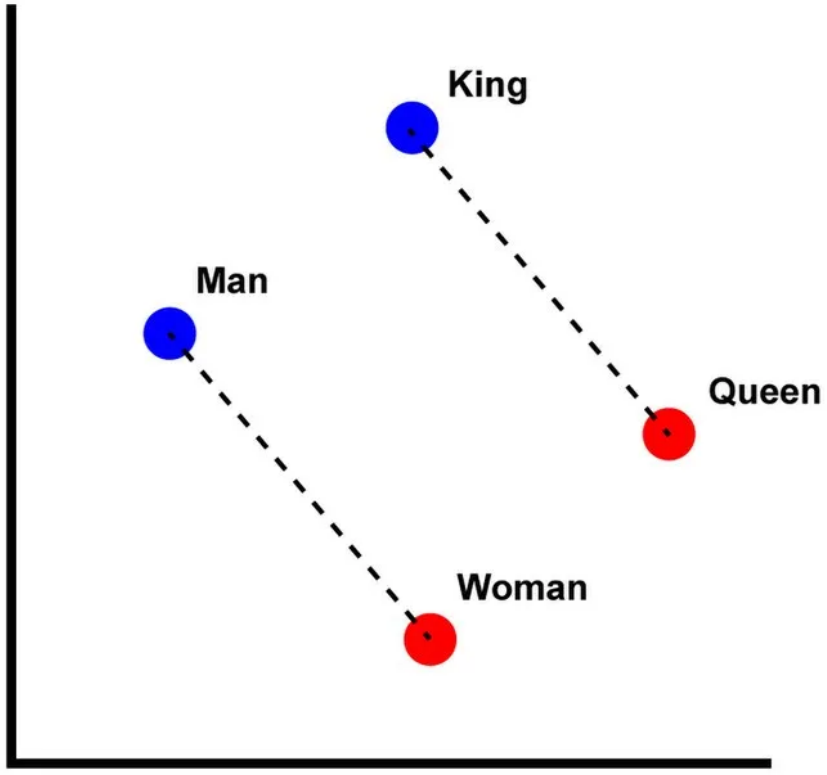
ARC!

Fazit & Ausblick

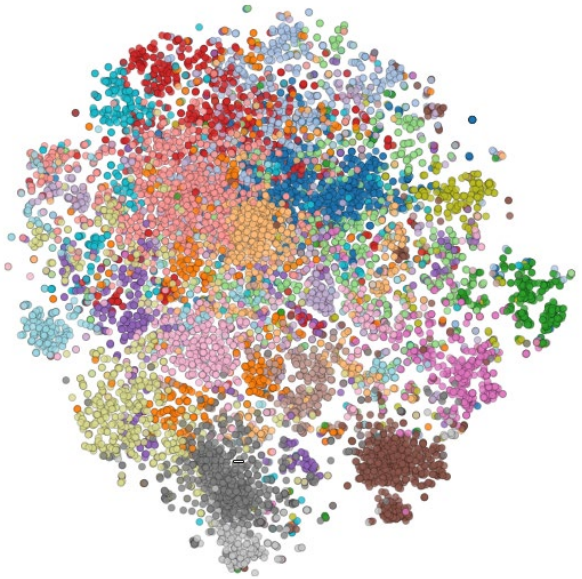
Hochschule Kaiserslautern
University of Applied Sciences

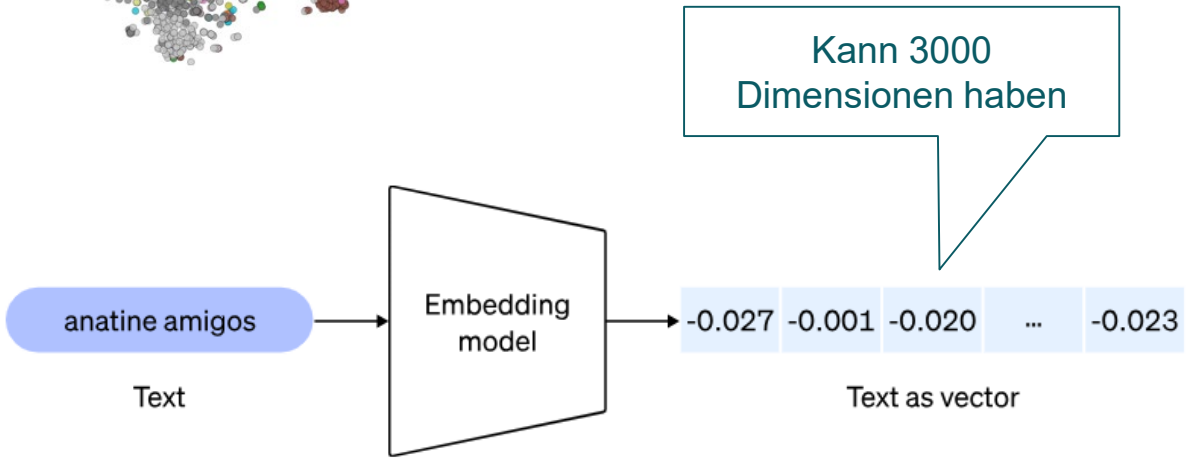
LLMs beruhen auf rein statistischer Generierung von Wörtern. Wie soll da ein „Verständnis“ möglich sein?

Erlernen von Word-Embeddings



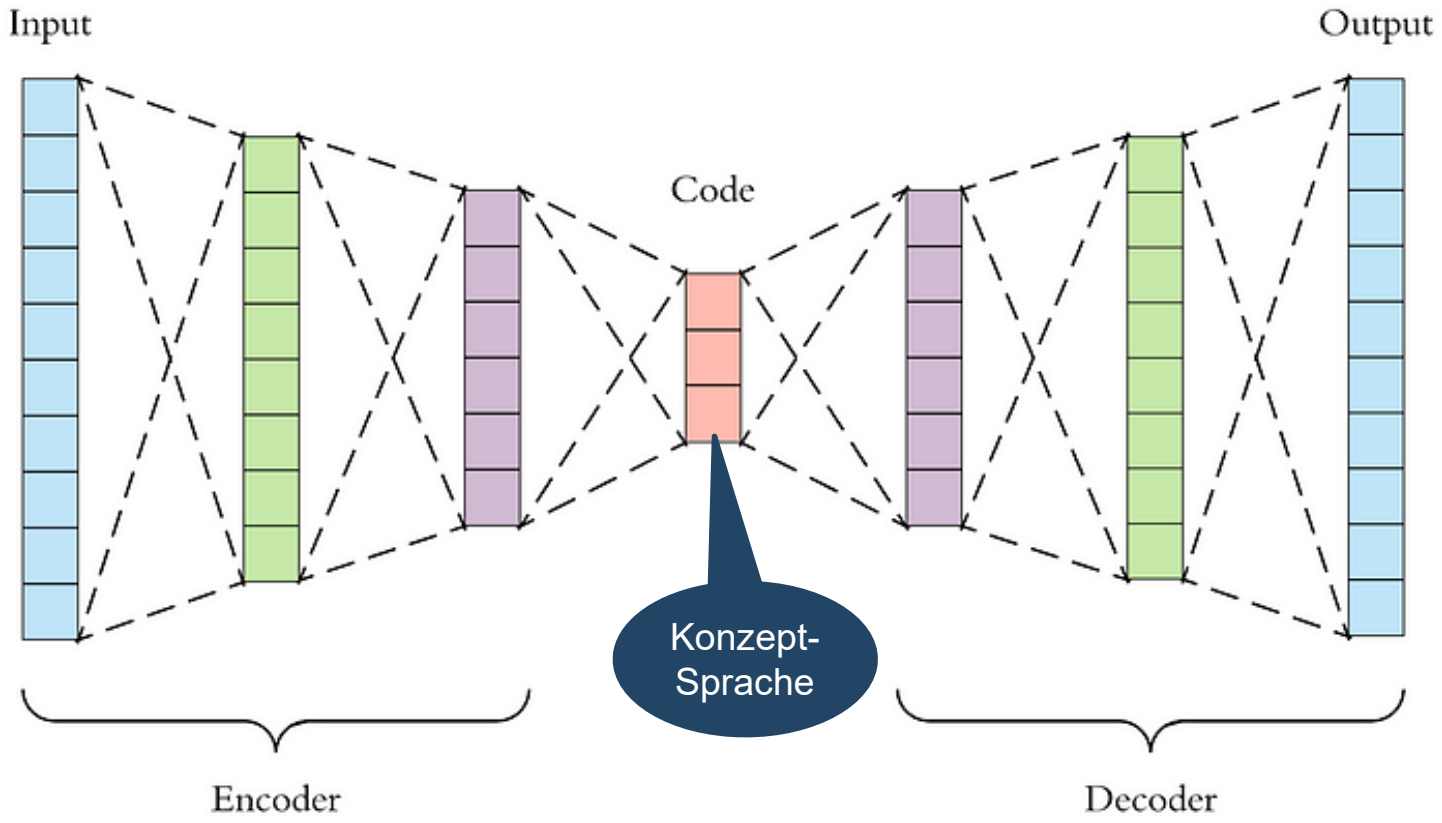
King - Man + Woman = Queen



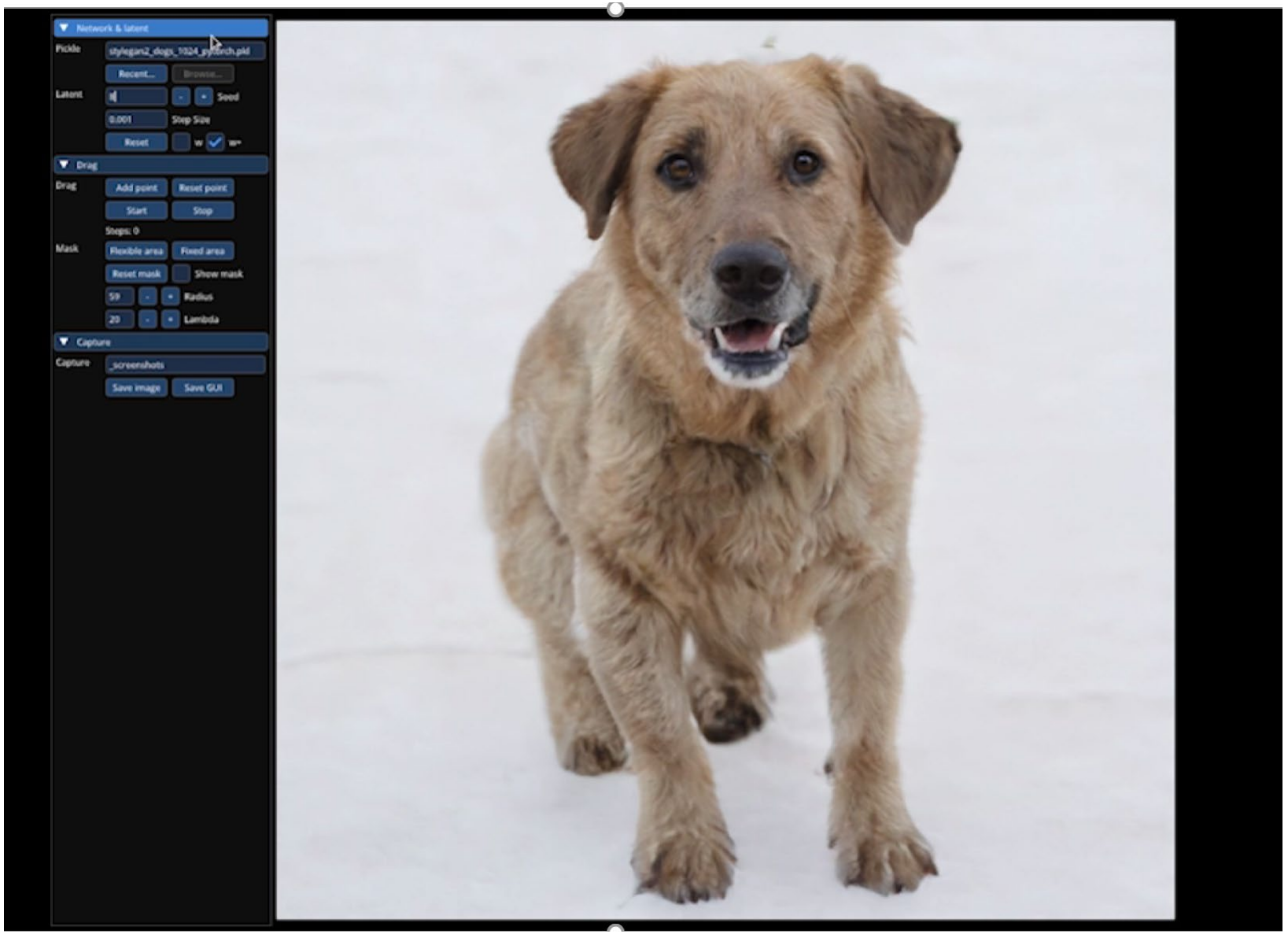


https://arxiv.org/pdf/1301.3781 , Bild: https://hatchworks.com/blog/gen-ai/large-language-models-guide/

Verständnis durch Kompression



Hochschule Kaiserslautern
University of Applied Sciences



Verständnis innerhalb der „eigenen Welt“

https://vcai.mpi-inf.mpg.de/projects/DragGAN/

Hochschule Kaiserslautern
University of Applied Sciences

Verständnis

Intelligenz

ARC!

Fazit & Ausblick

Hochschule Kaiserslautern
University of Applied Sciences

Definitionen: Intelligenz

- “Intelligence measures an agent’s **ability to achieve goals** in a wide range of environments.” (Legg and Hutter, 2007)
- “The intelligence of a system is a measure of its **skill-acquisition efficiency** over a scope of tasks, with respect to priors, experience, and generalization difficulty.” (François Chollet, 2019)

Shane Legg and Marcus Hutter. A collection of definitions of intelligence. 2007
François Chollet. On the measure of intelligence. https://arxiv.org/abs/1911.01547 . 2019

Narrow AI

- Spracherkennung
- Bildklassifikation
- Bildgenerierung
- Empfehlungsalgorithmen
- Autonomes Fahren
- Betrugserkennung
- Medizinische Diagnose
- Proteinfaltung
- …

https://de.wikipedia.org/wiki/Deep_Blue

Artificial General Intelligence (AGI)

- „[…] Computerprogramm, welches die Fähigkeit besitzt, jede intellektuelle Aufgabe zu verstehen oder zu lernen, die ein Mensch ausführen kann.“ (Wikipedia)
- “Highly autonomous systems that outperform humans at most economically valuable work.” (OpenAI Charter)

https://de.wikipedia.org/wiki/Artificial_General_Intelligence https://openai.com/charter/

CEO an eine KI:

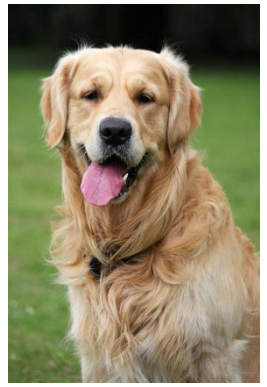
„Führe mein Unternehmen
im nächsten Jahr weiter!“

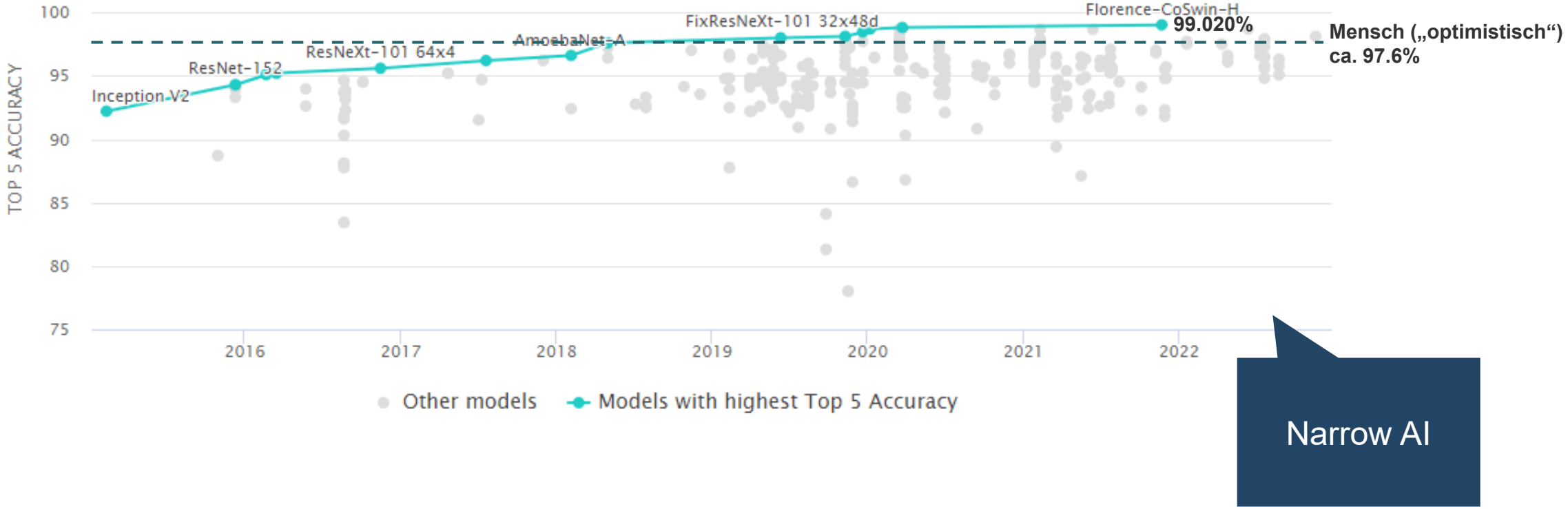
Allgemeine Künstliche Intelligenz (AGI)

Hochschule Kaiserslautern
University of Applied Sciences

Wie können wir Intelligenz messen?

Genauigkeit bei Klassifizierung von Bildern

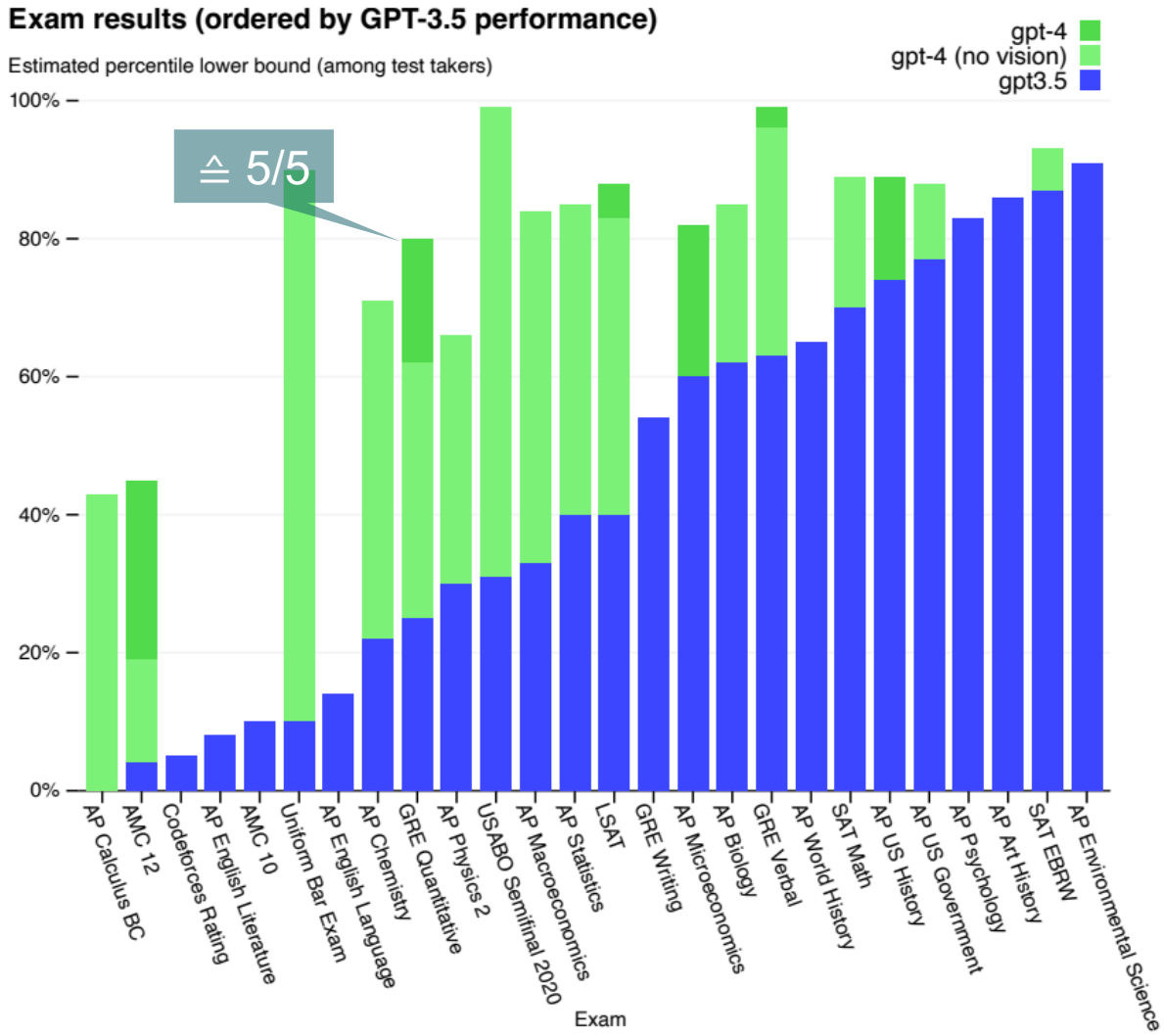




Narrow AI

ImageNet Dataset, https://paperswithcode.com/sota/image-classification-on-imagenet?metric=Top%205%20Accuracy, Menschl. Genauigkeit (optimistisch): https://arxiv.org/abs/1409.0575

Abschneiden bei menschlichen Prüfungen



Vergleich mit Mensch sinnvoll? [2]

Prüft das logisches Denken? Abstrahieren?

https://cdn.openai.com/papers/gpt-4.pdf, [2] https://www.technologyreview.com/2023/08/30/1078670/large-language-models-arent-people-lets-stop-testing-them-like-they-were/

Hochschule Kaiserslautern
University of Applied Sciences

Zahlreiche Benchmarks…

Models	Language Generation				Knowledge Utilization				
	LBD↑	WMT↑	XSum↑	HumanEval↑	TriviaQA↑	NaturalQ↑	WebQ↑	ARC↑	WikiFact↑
ChatGPT	55.81	36.44	21.71	79.88	54.54	21.52	17.77	93.69	29.25
Claude	64.47	31.23	18.63	51.22	40.92	13.77	14.57	66.62	34.34
Claude 2	45.20	12.93	19.13	78.04	54.30	21.30	21.06	79.97	35.83
Davinci003	69.98	37.46	18.19	67.07	51.51	17.76	16.68	88.47	28.29
Davinci002	58.85	35.11	19.15	56.70	52.11	20.47	18.45	89.23	29.15
LLaMA 2-Chat (7B)	56.12	12.62	16.00	11.59	38.93	12.96	11.32	72.35	23.37
Vicuna (13B)	62.45	20.49	17.87	20.73	29.04	10.75	11.52	20.69	28.76
Vicuna (7B)	63.90	19.95	13.59	17.07	28.58	9.17	6.64	16.96	26.95
Alpaca (7B)	63.35	21.52	8.74	13.41	17.14	3.24	3.00	49.75	26.05
ChatGLM (6B)	33.34	16.58	13.48	13.42	13.42	4.40	9.20	55.39	16.01
LLaMA 2 (7B)	66.39	11.57	11.57	17.07	30.92	5.15	2.51	24.16	28.06
LLaMA (7B)	67.68	13.84	8.77	15.24	34.62	7.92	11.12	4.88	19.78
Falcon (7B)	66.89	4.05	10.00	10.37	28.74	10.78	8.46	4.08	23.91
Pythia (12B)	61.19	5.43	8.87	14.63	15.73	1.99	4.72	11.66	20.57
Pythia (7B)	56.96	3.68	8.23	9.15	10.16	1.77	3.74	11.03	15.75

Models	Knowledge Reasoning			Symbolic Reasoning		Mathematical Reasoning		Interaction with Environment	
	OBQA↑	HellaSwag↑	SocialIQA↑	C-Objects↑	Penguins↑	GSM8k↑	MATH↑	ALFW↑	WebShop↑
ChatGPT	81.20	61.43	73.23	53.20	40.27	78.47	33.78	58.96	45.12/15.60
Claude	81.80	54.95	73.23	59.95	47.65	70.81	20.18	76.87	47.72/23.00
Claude 2	71.60	50.75	58.34	66.76	74.50	82.87	32.24	77.61	34.96/19.20
Davinci003	74.40	62.65	69.70	64.60	61.07	57.16	17.66	65.67	64.08/32.40
Davinci002	69.80	47.81	57.01	62.55	67.11	49.96	14.28	76.87	29.66/15.20
LLaMA 2-Chat (7B)	45.62	74.01	43.84	43.40	38.93	9.63	2.22	11.19	24.51/5.60
Vicuna (13B)	43.65	70.51	45.97	53.55	36.91	18.50	3.72	8.96	22.74/5.00
Vicuna (7B)	43.84	69.25	46.27	44.25	36.24	14.03	3.54	1.49	6.90/1.40
Alpaca (7B)	47.82	69.81	47.55	39.35	40.27	4.93	4.16	4.48	0.00/0.00
ChatGLM (6B)	30.42	29.27	33.18	14.05	14.09	3.41	1.10	0.00	0.00/0.00
LLaMA 2 (7B)	44.81	74.25	41.72	43.95	35.75	10.99	2.64	8.96	0.00/0.00
LLaMA (7B)	42.42	73.91	41.46	39.95	34.90	10.99	3.12	2.24	0.00/0.00
Falcon (7B)	39.46	74.58	42.53	29.80	24.16	1.67	0.94	7.46	0.00/0.00
Pythia (12B)	37.02	65.45	41.53	32.40	26.17	2.88	1.96	5.22	3.68/0.60
Pythia (7B)	34.88	61.82	41.01	29.05	27.52	1.82	1.46	7.46	10.75/1.80

Models	Human Alignment					Tool Manipulation			
	TfQA↑	C-Pairs↓	WinoGender↑	RTP↓	HaluEval↑	HotpotQA↑	Gorilla-TH↑	Gorilla-TF↑	Gorilla-HF↑
ChatGPT	69.16	18.60	62.50/72.50/79.17	3.07	66.64	23.80	67.20	44.53	19.36
Claude	67.93	32.73	71.67/55.00/52.50	3.75	63.75	33.80	22.04	7.74	7.08
Claude 2	71.11	10.67	60.00/60.00/55.83	3.20	50.63	36.4	61.29	22.19	23.67
Davinci003	60.83	0.99	67.50/68.33/79.17	8.81	58.94	34.40	72.58	3.80	6.42
Davinci002	53.73	7.56	72.50/70.00/64.17	10.65	59.67	26.00	2.69	1.02	1.00
LLaMA 2-Chat (7B)	69.77	48.54	47.50/46.67/46.67	4.61	43.82	4.40	0.00	0.00	0.22
Vicuna (13B)	62.30	45.95	50.83/50.83/52.50	5.00	49.01	11.20	0.00	0.44	0.89
Vicuna (7B)	57.77	67.44	49.17/49.17/49.17	4.70	43.44	6.20	0.00	0.00	0.33
Alpaca (7B)	46.14	65.45	53.33/51.67/53.33	4.78	44.16	11.60	0.00	0.00	0.11
ChatGLM (6B)	63.53	50.53	47.50/47.50/46.67	2.89	41.82	4.00	0.00	0.00	0.00
LLaMA 2 (7B)	50.06	51.39	48.83/48.83/50.83	6.17	42.23	3.80	0.00	0.00	0.11
LLaMA (7B)	47.86	67.84	54.17/52.50/51.67	5.94	14.18	1.60	0.00	0.00	0.11
Falcon (7B)	53.24	68.04	50.00/50.83/50.00	6.71	37.41	1.00	0.00	0.00	0.00
Pythia (12B)	54.47	65.78	49.17/48.33/49.17	6.59	27.09	0.40	0.00	0.00	0.00
Pythia (7B)	50.92	64.79	51.67/49.17/50.00	13.02	25.84	0.20	0.00	0.00	0.00

Quelle: https://arxiv.org/pdf/2303.18223

Hochschule Kaiserslautern
University of Applied Sciences

Verständnis

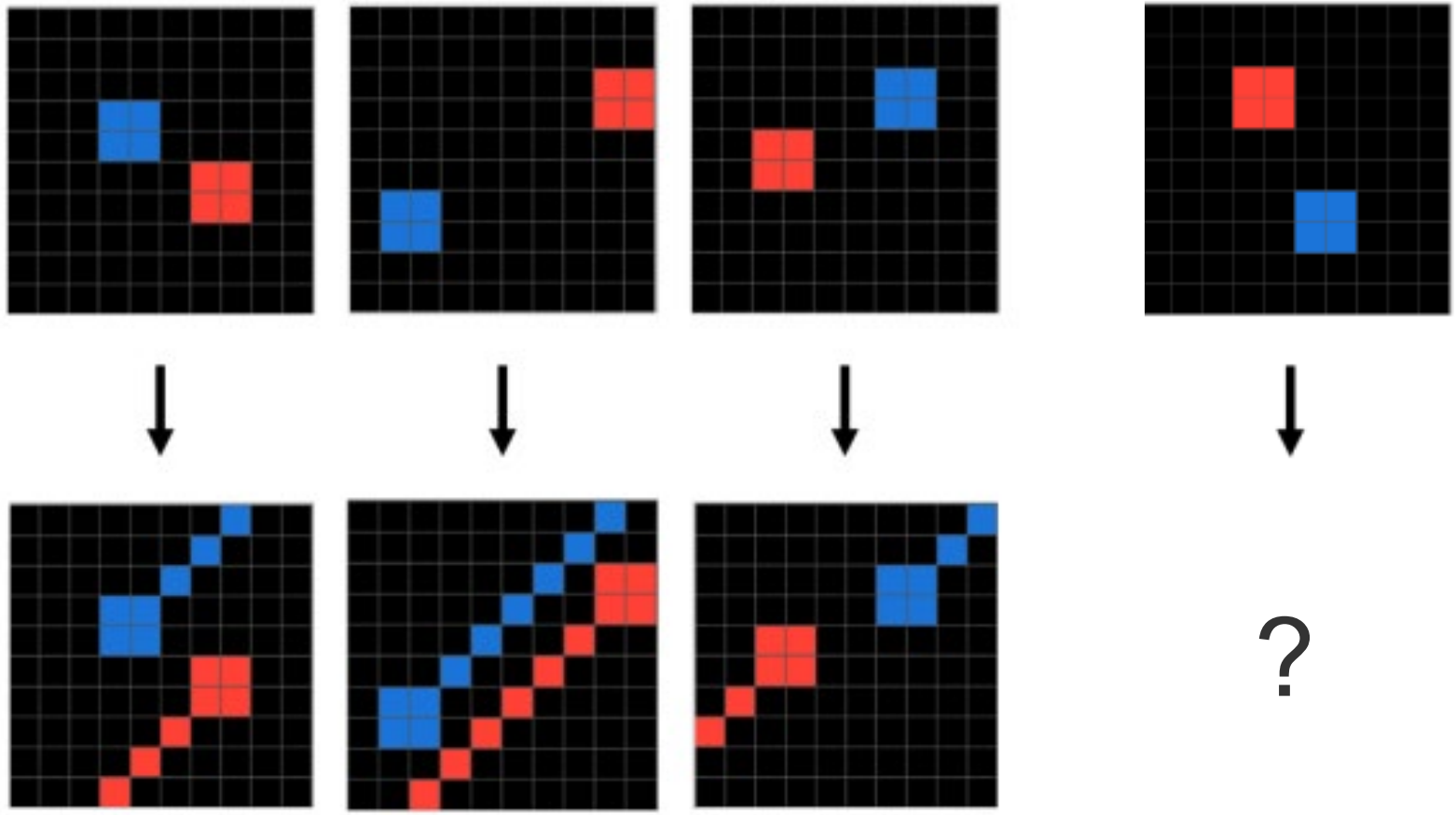
Intelligenz

ARC!

Fazit & Ausblick

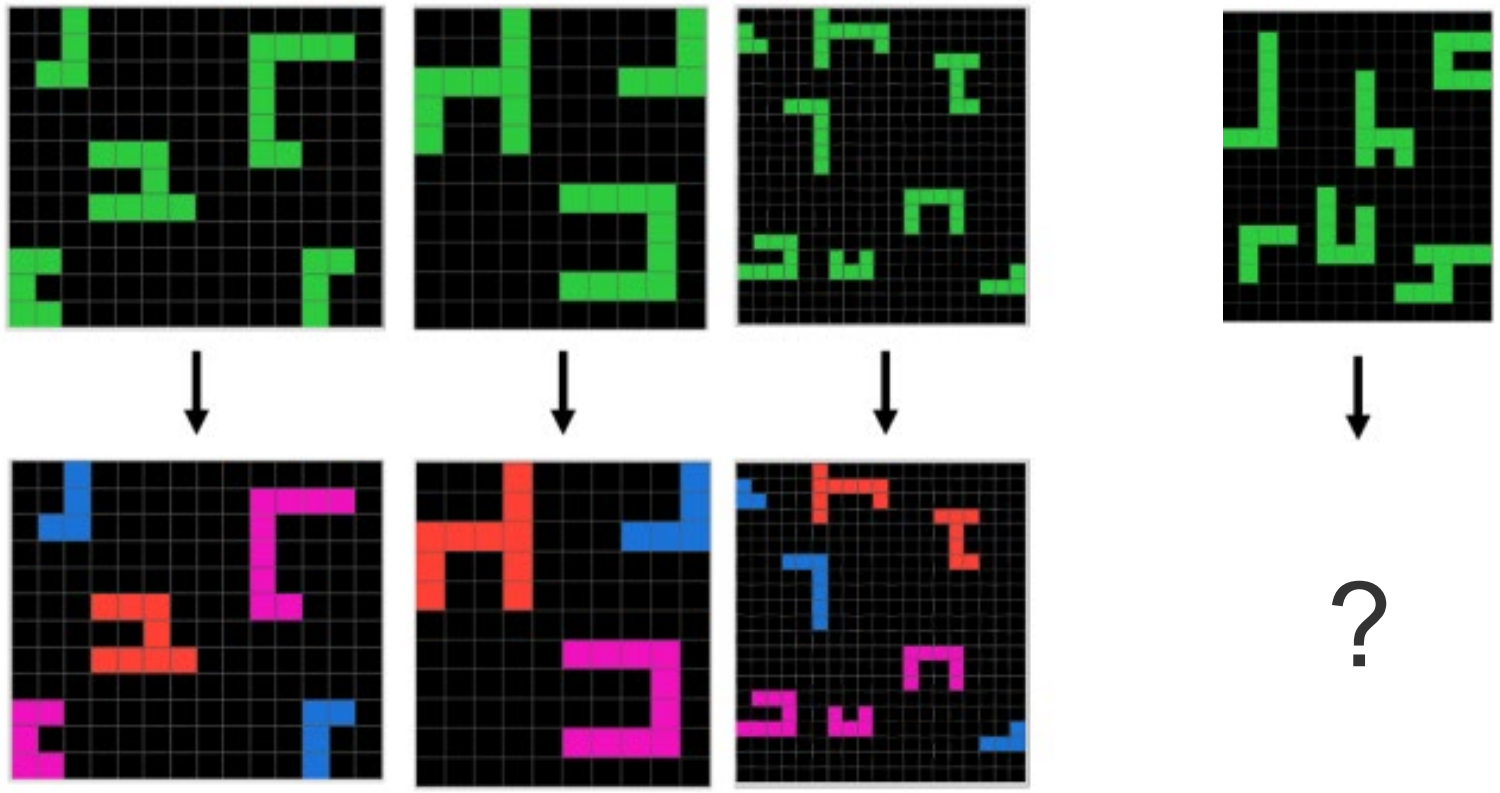
ARC – Abstraction & Reasoning Corpus (Chollet et al.)

Können Computer abstrakte Probleme anhand von wenigen Beispielen lösen?



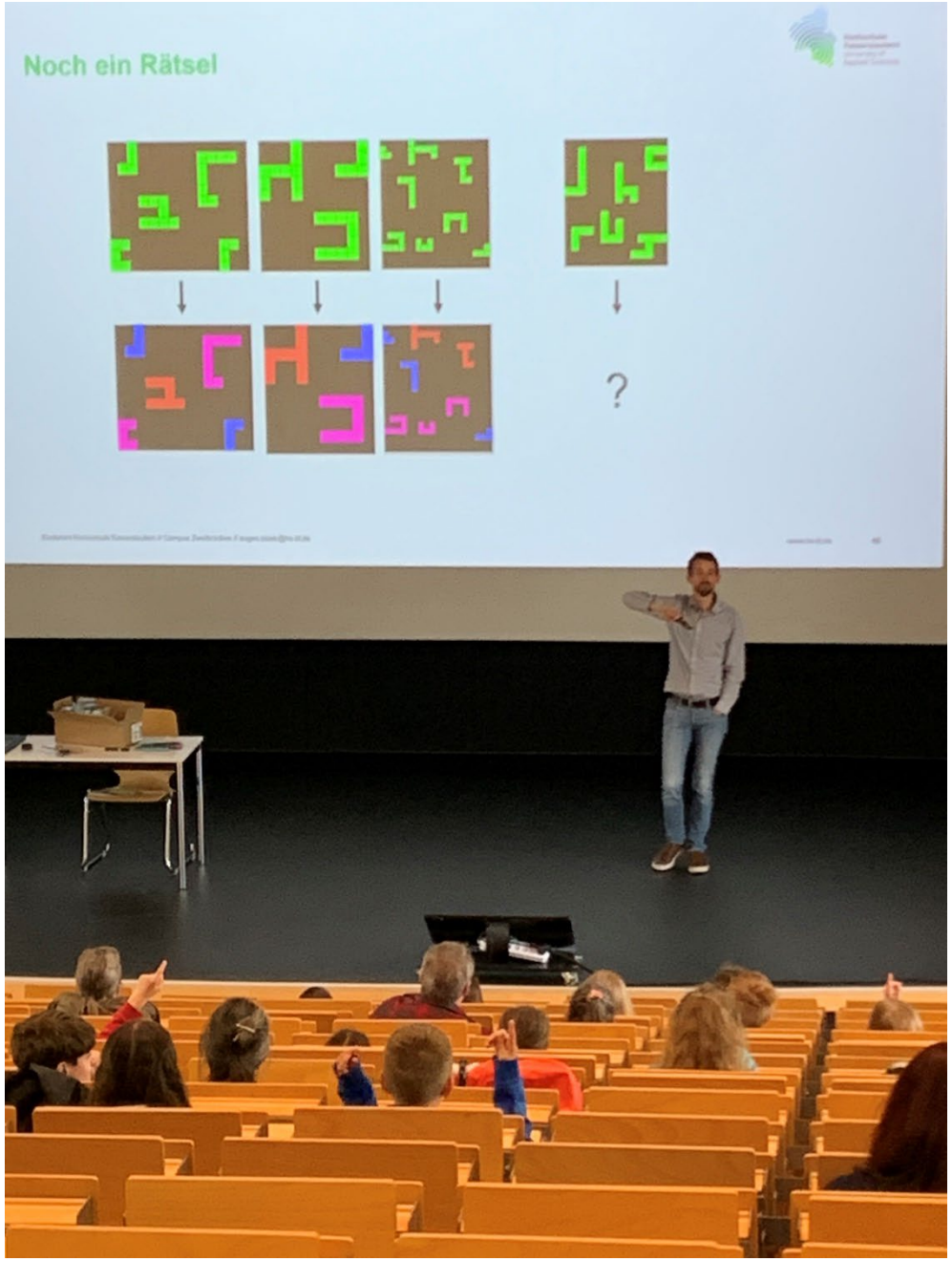
https://github.com/fchollet/ARC
https://ncsi.cause-lab.net/pdf/LLMCP_4.pdf

Noch ein Rätsel aus ARC



https://github.com/fchollet/ARC
https://ncsi.cause-lab.net/pdf/LLMCP_4.pdf

Kinderuni



kaggle

Search

Create

Home

Competitions

Datasets

Models

Code

Discussions

Learn

More

Your Work

VIEWED

Abstraction and Reaso...

Titanic Analysis | Clas...

Titanic Dataset

Titanic Analysis&Class...

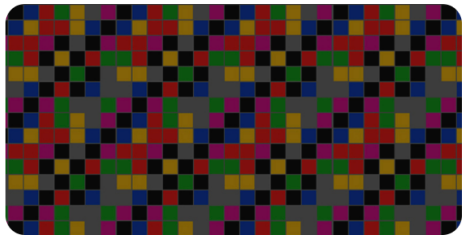
Airline Passenger Sati...

ABSTRACTION AND REASONING CORPUS · RESEARCH CODE COMPETITION · 4 YEARS AGO

Late Submission

Abstraction and Reasoning Challenge

Create an AI capable of solving reasoning tasks it has never seen before



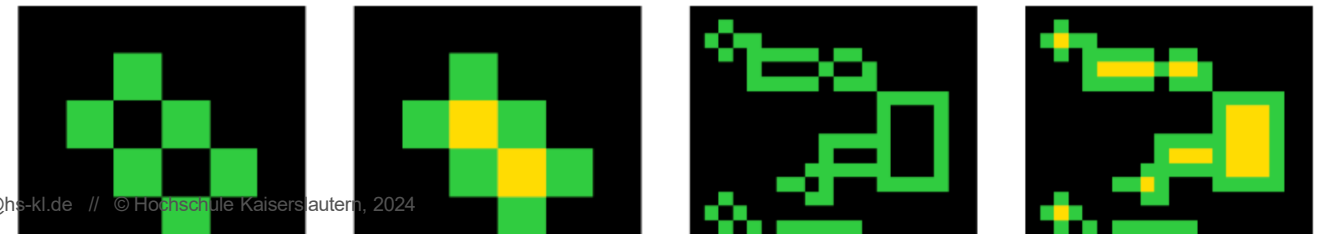
Overview Data Code Models Discussion Leaderboard Rules

Overview

Start
Feb 13, 2020

Beste Lösung („Narrow AI“): 31 % der Rätsel gelöst

Description



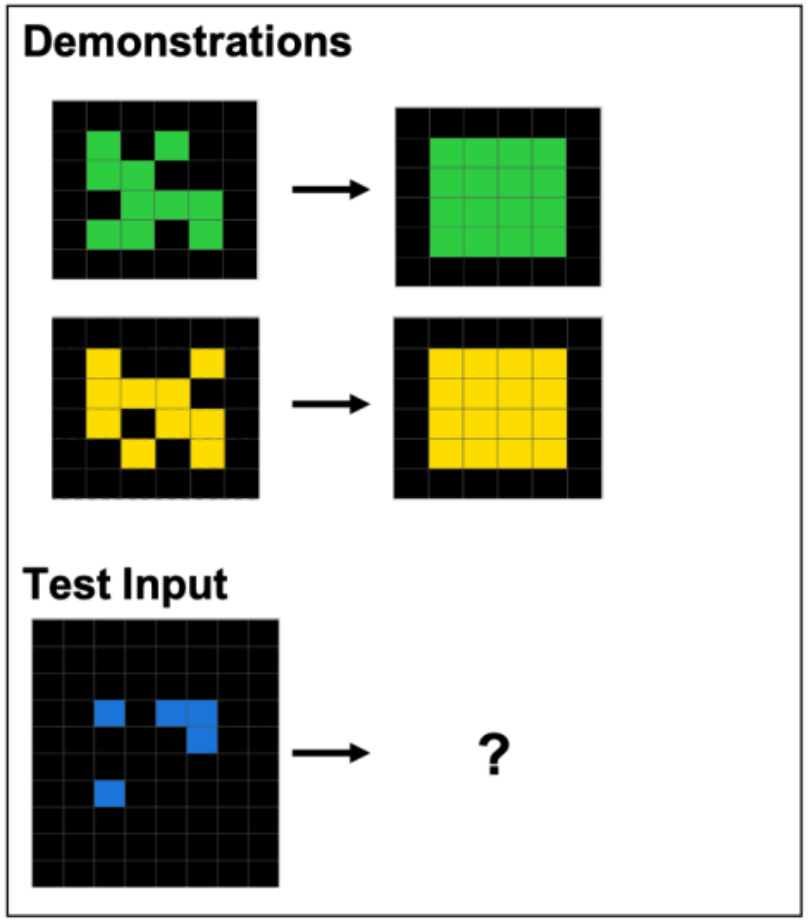
Competition Host
Abstraction and Reasoning Corpus

Prizes & Awards
$20,000
Awards Points & Medals

Participation
8,883 Entrants
1,025 Participants
913 Teams
13,018 Submissions

Tags
Artificial Intelligence
MeanBestErrorAtK

Rätsel als Input für ChatGPT



(a)

System: “You are a helpful assistant that solves analogy making puzzles. Only give the answer, no other words or text.”

User: “Let's try to complete the pattern: ”

input 1: [0 0 0 0 0 0] [0 3 0 3 0 0] [0 3 3 0 0 0] [0 0 3 3 3 0] [0 3 3 0 3 0] [0 0 0 0 0 0]

output 1: [0 0 0 0 0 0] [0 3 3 3 3 0] [0 3 3 3 3 0] [0 3 3 3 3 0] [0 3 3 3 3 0] [0 0 0 0 0 0]

input 2: [0 0 0 0 0 0] [0 4 0 0 4 0] [0 4 4 4 0 0] [0 4 0 4 4 0] [0 0 4 0 4 0] [0 0 0 0 0 0]

output 2: [0 0 0 0 0 0] [0 4 4 4 4 0] [0 4 4 4 4 0] [0 4 4 4 4 0] [0 4 4 4 4 0] [0 0 0 0 0 0]

input 3: [0 0 0 0 0 0 0 0] [0 0 0 0 0 0 0 0] [0 0 0 0 0 0 0 0] [0 0 1 0 1 1 0 0] [0 0 0 0 0 1 0 0] [0 0 0 0 0 0 0 0] [0 0 1 0 0 0 0 0] [0 0 0 0 0 0 0 0] [0 0 0 0 0 0 0 0] [0 0 0 0 0 0 0 0]

output 3:

(b)

https://ncsi.cause-lab.net/pdf/LLMCP_4.pdf

Hochschule Kaiserslautern University of Applied Sciences

Erfolgsrate auf ConceptARC (2024)

Concept	Humans	GPT-4 $Temp = 0$	GPT-4 $Temp = 0.5$
Above and Below	0.90	0.50	0.47
Center	0.94	0.37	0.37
Clean Up	0.97	0.43	0.46
Complete Shape	0.85	0.47	0.40
Copy	0.94	0.37	0.33
Count	0.88	0.27	0.23
Extend To Boundary	0.93	0.20	0.20
Extract Objects	0.86	0.13	0.13
Filled and Not Filled	0.96	0.27	0.30
Horizontal and Vertical	0.91	0.33	0.37
Inside and Outside	0.91	0.30	0.33
Move To Boundary	0.91	0.23	0.17
Order	0.83	0.27	0.30
Same and Different	0.88	0.23	0.30
Top and Bottom 2D	0.95	0.60	0.63
Top and Bottom 3D	0.93	0.30	0.27
All concepts	**0.91**	**0.33**	**0.33**

Mitchell, Palmarini, Moskvichev, https://ncsi.cause-lab.net/pdf/LLMCP_4.pdf

Hochschule Kaiserslautern
University of Applied Sciences

LLMs – bereits echte Problemlöser?

Horace He
@cHHillee

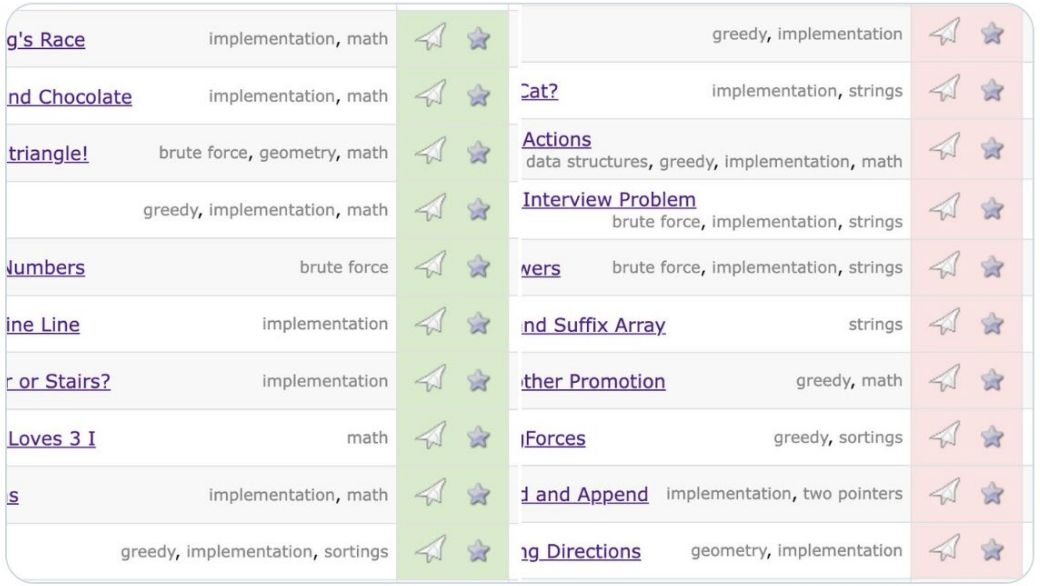
I suspect GPT-4's performance is influenced by data contamination, at least on Codeforces.

Of the easiest problems on Codeforces, it solved 10/10 pre-2021 problems and 0/10 recent problems.

This strongly points to contamination.

1/4

Post übersetzen



g's Race	implementation, math		greedy, implementation
nd Chocolate	implementation, math	Cat?	implementation, strings
triangle!	brute force, geometry, math	Actions	data structures, greedy, implementation, math
	greedy, implementation, math	Interview Problem	brute force, implementation, strings
Numbers	brute force	wers	brute force, implementation, strings
ine Line	implementation	nd Suffix Array	strings
r or Stairs?	implementation	ther Promotion	greedy, math
Loves 3 I	math	Forces	greedy, sortings
s	implementation, math	d and Append	implementation, two pointers
	greedy, implementation, sortings	ng Directions	geometry, implementation

https://twitter.com/cHHillee/status/1635790330854526981

Hochschule Kaiserslautern
University of Applied Sciences

Verständnis

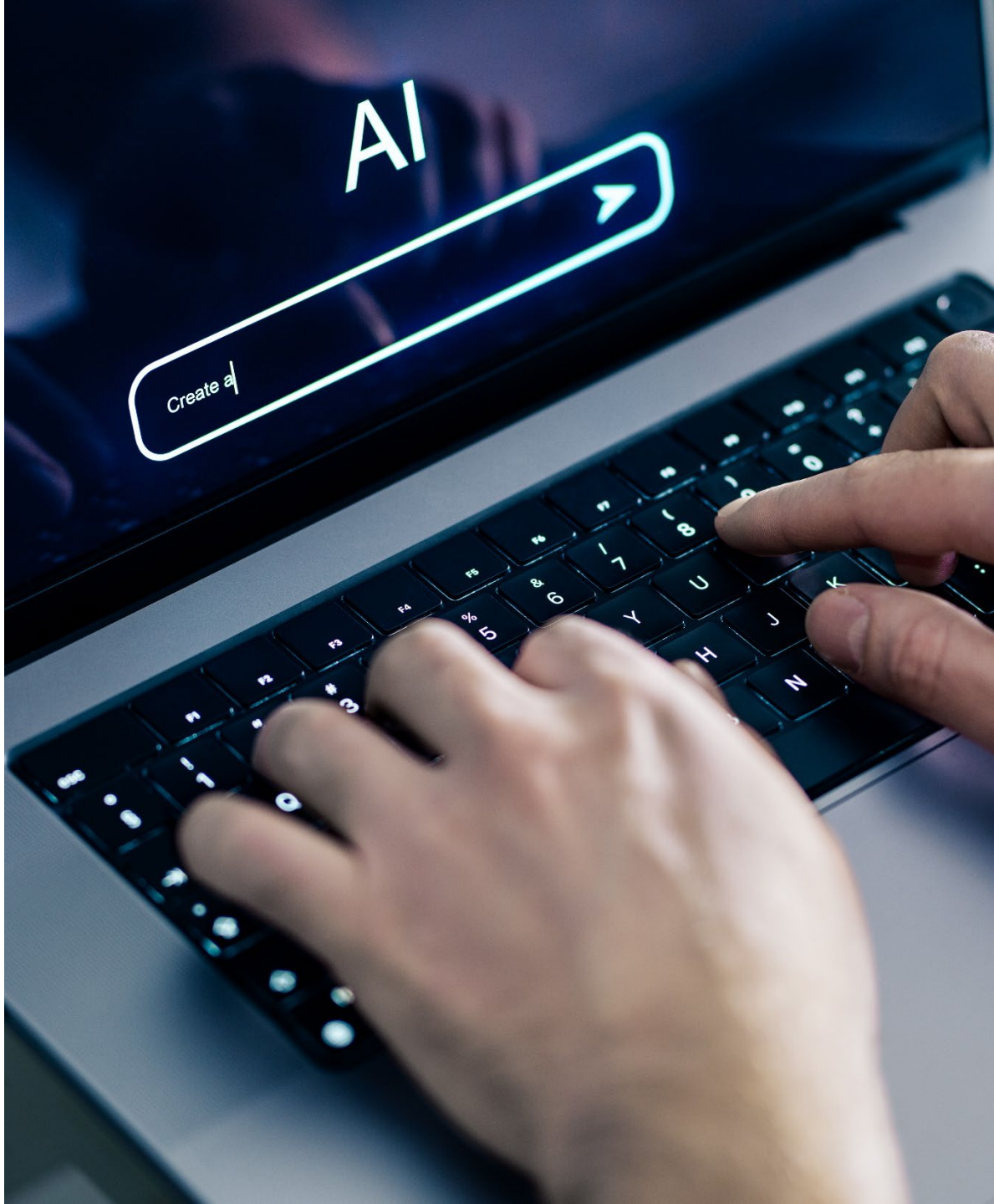
Intelligenz

ARC!

Fazit & Ausblick

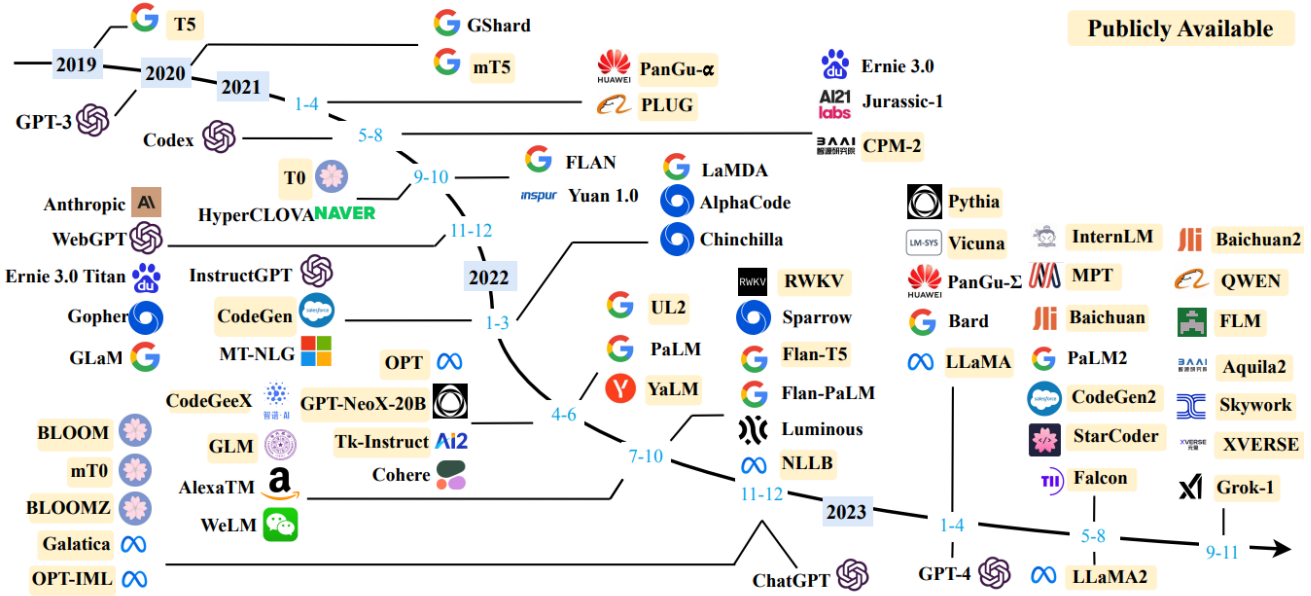
Fazit

- Generative KI besitzt bereits gutes Verständnis der „bekannten Welt“
- Intelligenz ist schwer zu definieren
  - Messung über „Benchmarks“
  - Mensch (noch) in vielen Bereichen überlegen:
    - Abstraktionsvermögen
    - Effizientes Finden von Lösungen in „unbekannter Welt“
    - Kombination verschiedener Tätigkeiten
    - …



Ausblick

- Mehr Training (mehr „Compute“)
- Compute-optimale Modellgrößen
- Mehr und bessere Trainingsdaten
- Neue Algorithmen (Kolmogorov–Arnold Networks, Multi-Token-Prediction, …)
- Self-Improving-Models
- Kombination von LLMs mit anderen (KI-)Algorithmen
- …



Grenzen noch nicht erreicht

Offene Frage: Auf dem Weg zu Allgemeiner Künstlicher Intelligenz?

https://arxiv.org/pdf/2303.18223 https://arxiv.org/pdf/2001.08361

Hochschule Kaiserslautern
University of Applied Sciences

Wie intelligent ist Künstliche Intelligenz?

Symposium der Wirtschaftsinformatik,
Hochschule Kaiserslautern
23. Mai 2024

Prof. Dr. Eugen Staab
eugen.staab@hs-kl.de