Ralf Walther
Geschäftsführer & Gründer
mindUp Web + Intelligence GmbH

mindUp
Web + Intelligence GmbH

Vorstand & Gründungsmitglied
cyberLAGO e.V.

cyberLAGO
digital competence network

KI-LAB Bodensee
KI-Kompetenz für die Region

KI-Kompetenz aufbauen

KI-Anwendungen erproben

KI-Experten finden

Was gibt‘s denn heute?

Ist das Spielerei oder kann das weg?

European Centre for Medium-Range Weather Forecasts

ECMWF | 50

Search site...

Help

Log in

Home About Forecasts Computing Research Learning

Publications Anniversary

Who we are | What we do | Jobs | **Media centre** | Suppliers | Location

First AIFS model weights are now open

11 December 2024
Mario Santa Cruz
Gert Mertes
The AIFS team

Ana Prieto Nemesio
Baudouin Raoult

◀ View all AIFS blog posts

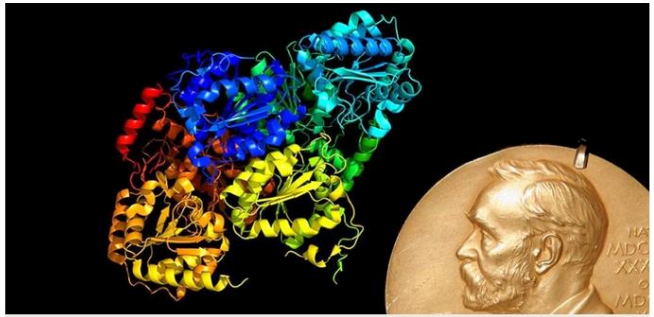
We are excited to announce a key step in making our data-driven weather model even more accessible to a broader audience of developers, scientists, and weather enthusiasts. Already you can access real-time charts and data for the AIFS, and now the latest deterministic version of the AIFS (AIFS-single), v0.2.1, is available on Hugging Face so you can run the AIFS on your own device. Hugging Face is a widely used platform and community hub for sharing machine learning models. It provides an open ecosystem where developers can share, access, and experiment with state-of-the-art models across disciplines. By hosting

Chemie

Chemie-Nobelpreis für KI-gestützte Proteinforschung

Entwickler der KI AlphaFold und Pionier künstlicher Proteine ausgezeichnet

9. Oktober 2024, Lesezeit: 4 Min.



Den Chemie-Nobelpreis 2024 erhalten drei Pioniere der computergestützten Proteinforschung. © gemeinfrei

Product Solutions Resources Open Source

google-deepmind / alphafold3 Public

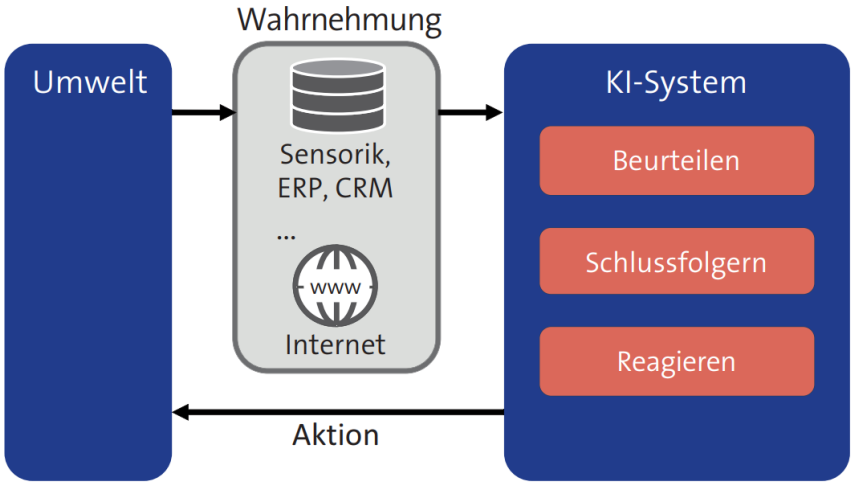
AlphaFold 3

This package provides an implementation of the inference pipeline of AlphaFold 3. See below for how to access the model parameters. You may only use AlphaFold 3 model parameters if received directly from Google. Use is subject to these terms of use.

Any publication that discloses findings arising from using this source code, the model parameters or outputs produced by those should cite the Accurate structure prediction of biomolecular interactions with AlphaFold 3 paper.

Please also refer to the Supplementary Information for a detailed description of the method.

AlphaFold 3 is also available at alphafoldserver.com for non-commercial use, though with a more limited set of ligands and covalent modifications.



Was nehme ich wahr?

Was kann ich ableiten?

Wie muss ich reagieren?

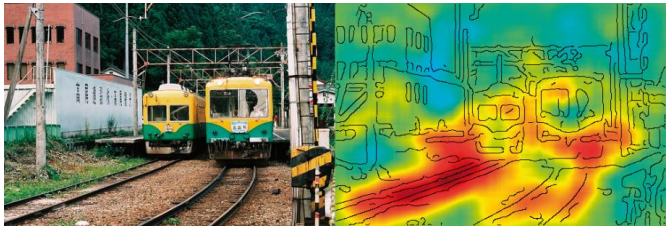
Quelle: www.igcv.fraunhofer.de/de/forschung/kompetenzen/kuenstliche_intelligenz.html

Was soll mit dem Einsatz von KI erreicht werden?

- Effizienzsteigerung
  - Automatisierung von wiederkehrenden Aufgaben
  - Analyse von großen Datenmengen
  - Konzentration der Akteure auf komplexe Aufgaben, die (noch) nicht von KI gelöst werden können
- Qualitätsverbesserung
  - Validierung/Standardisierung von Ergebnissen
  - Messbarkeit
- Kommunikation/Kollaboration

KI, was kann da schon schiefgehen?

- Überhöhte Erwartungen, unterschätzte Unschärfe
- Falsche Zieldefinition aufgrund Unkenntnis der Möglichkeiten
- Zu wenig oder ungeeignete Daten: Ohne Daten kann keine KI-basierte Lösung erfolgreich sein
- Was nicht in den Daten steckt, kann nicht von der KI gelernt werden
- Interpretierbarkeit des Modells/der Ergebnisse



Quelle: https://www.hhi.fraunhofer.de/presse-medien/nachrichten/2019/wie-intelligent-ist-kuenstliche-intelligenz.html

KI weiß meistens nicht, was es nicht weiß

Realistische Darstellung von mehreren Armbanduhren auf einem modernen Tisch, die alle die Zeit 17:35 zeigen.

Realistische Darstellung eines Fortbildungskurses für Zeichnen. An einem Tisch sitzen mehrere Linkshänder, die mit der linken Hand eine Zeichnung anfertigen.

Finde den Fehler!

Der Takt wird schneller

ChatGPT / GPT-4
2022

OpenAI

23.07.24
Llama 3.1

04.03.24
ANTHROP\C
Claude

27.03.24
DBRX
databricks

27.03.24
Amazon spends $2.75 billion on AI startup Anthropic in its largest venture investment yet
CNBC

13.08.24
Grok-2 Beta Release

12.09.24
OpenAI o1

22.10.24
Teaching Claude to navigate computers, responsibly
Claude 3.5 Sonnet
ANTHROP\C

06.12.24
Llama 3.3
Meta

11.12.24
DeepMind
Gemini 2.0

13.12.24
MICROSOFT
Phi-4

20.01.25
deepseek
R1

19.02.25
GROK
Grok 3 Beta — The Age of Reasoning Agents

18.02.25
perplexity
Today we're open-sourcing R1 1776, a version of the DeepSeek-R1 model that has been post-trained to provide unbiased, accurate, and factual information.

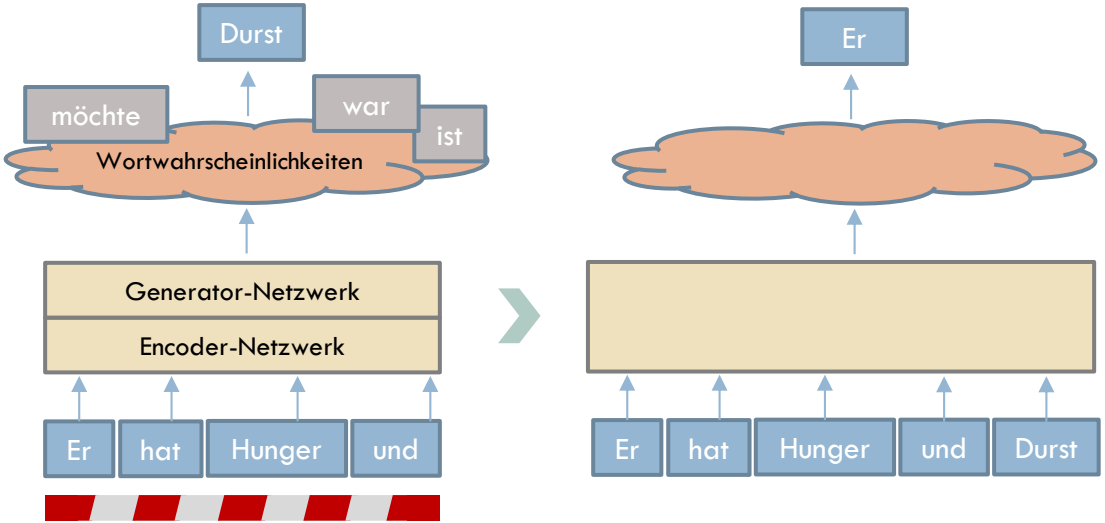
24.02.25
Claude 3.7 Sonnet and Claude Code

12.03.25
google
Gemma 3

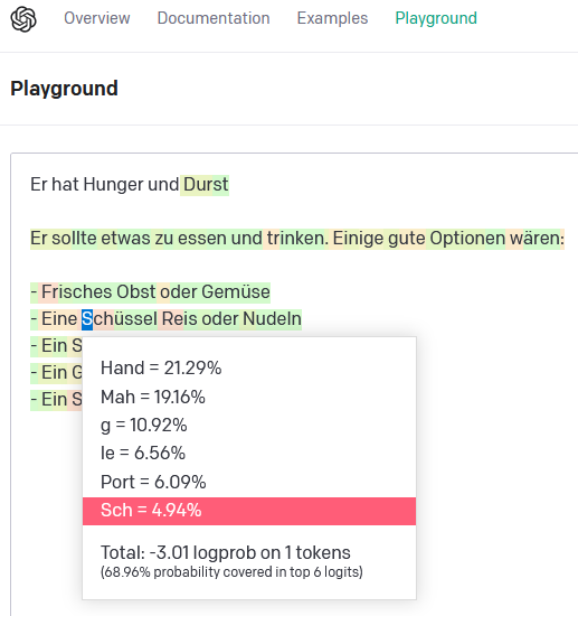
Model	Lab	Playground	Parameters (B)	Tokens trained (B)	Ratio Tokens:Params (Chinchilla scaling≥20:1)	ALScore "ALScore" is Sqr Root of	MMLU	MMLU -Pro	GPQA	Training dataset	Announced ▼	Public?	Paper / Repo	Arch	Tags	Notes
DeepSeek-V3 0324	DeepSeek-AI	https://chat.deep	685	14800	22:1	10.6		81.2	68.4	synthetic, web-scale	Mar/2025		https://hug	MoE	SOTA	Non-reasoning. Significant increase in benchmark performance compared to original V3
Llama-3.3-Nemotron	NVIDIA	https://huggingfa	49	*15040*	307:1	2.9			66.67	web-scale	Mar/2025		https://built	Dense	Reasoning	Meta Llama-3.3-70B-Instruct derivative "that is post trained for reasoning, human chat p
EXAONE Deep	LG	https://huggingfa	32	6500	204:1	1.5	83	74	66.1	web-scale	Mar/2025		https://arxiv	Dense	Reasoning	"EXAONE"="EXpert AI for EveryONE". Training tokens/ratio dropped from EXAONE-3 7.8
Mistral Small 3.1	Mistral	https://huggingfa	24	*8000*	334:1	1.5	81.01	56.03	37.5	web-scale	Mar/2025		https://mist	Dense		"Mistral Small 3.1 (2503) adds state-of-the-art vision understanding and enhances long c
ERNIE 4.5	Baidu	https://yiyan.baic						*79*	*55*	synthetic, web-scale	Mar/2025		https://ww	*MoE*		Announce: https://x.com/Baidu_Inc/status/1901094083508220035
X1	Baidu	https://yiyan.baic								synthetic, web-scale	Mar/2025		https://ww	*MoE*	Reasoning	
OLMo 2 32B	Allen AI	https://playgrour	32	6400	200:1	1.5	78			synthetic, web-scale	Mar/2025		https://aller	Dense		"the first fully-open model (all data, code, weights, and details are freely available) to ou
Command A	Cohere	https://dashboar	111	*8000*	73:1	3.1	85			synthetic, web-scale	Mar/2025		https://hug	Dense		Context=256k. "Command A is an open weights research release of a 111 billion parame
Gemini Robotics	Google DeepMind		*200*	*20000*	100:1	6.7		*79.1*	*64.7*	synthetic, web-scale	Mar/2025		https://stor	*MoE*		Gemini 2.0 Pro (cloud). "The second model is Gemini Robotics, a state-of-theart Vision-L
Gemini Robotics-ER	Google DeepMind		*30*	*30000*	1,000:1	3.2	*87*	*76.4*	*62.1*	synthetic, web-scale	Mar/2025		https://stor	*MoE*		Gemini 2.0 Flash (on device). "The first model is Gemini Robotics-ER, a VLM with strong
Gemma 3	Google DeepMind	https://huggingfa	27	14000	519:1	2.0	78.6	67.5	42.4	web-scale	Mar/2025		https://stor	Dense		Trained on 1T more tokens than Gemma 2. "introduces vision understanding abilities, a
Reka Flash 3	Reka AI	https://huggingfa	21	*5000*	239:1	1.1		65	61.1	web-scale	Mar/2025		https://ww	Dense	Reasoning	"performs competitively with proprietary models such as OpenAI o1-mini, making it a g
QwQ-32B	Alibaba	https://huggingfa	32	18000	563:1	2.5				synthetic, web-scale	Mar/2025		https://qwe	Dense	Reasoning	Update to QwQ-32B-Preview released Nov/2024. Scores 1/5 on latest ALPrompt 2024 H
Jamba 1.6	AI21	https://huggingfa	398	*8000*	21:1	5.9	*81.2*	*53.5*	*36.9*	web-scale	Mar/2025		https://hug	MoE		"The AI21 Jamba 1.6 family of models is state-of-the-art, hybrid SSM-Transformer instru
Instella-3B	AMD	https://huggingfa	3	4160	1,387:1	0.4	58.31		30.13	web-scale	Mar/2025		https://rocr	Dense		"trained from scratch on AMD Instinct™ MI300X GPUs. Instella models outperform exist
Babel-83B	Alibaba	https://huggingfa	83	*15000*	181:1	3.7				synthetic, web-scale	Mar/2025		https://arxiv	Dense		"top 25 languages by number of speakers, including English, Chinese, Hindi, Spanish, Ara
Granite-3.2-8B-Instr	IBM	https://huggingfa	8	*12000*	1,500:1	1.0	66.79			web-scale	Feb/2025		https://ww	Dense	Reasoning	"The new Granite 3.2 8B Instruct [offers] experimental chain-of-thought reasoning capal
C4AI Command R7B	Cohere	https://huggingfa	7	*2000*	286:1	0.4		29.4	7.9	web-scale	Feb/2025		https://hug	Dense		"C4AI Command R7B Arabic is an open weights research release of a 7 billion parameter
GPT-4.5	OpenAI	https://chat.com/	*5400*	*114000*	22:1	82.7	89.6		71.4	synthetic, web-scale	Feb/2025		https://ope	MoE	SOTA	"Our largest and best model for chat" https://openai.com/index/introducing-gpt-4-5/ "G
Hunyuan T1	Tencent	https://cloud.ten	*389*	*7000*	18:1	5.5		87.2	69.3	synthetic, web-scale	Feb/2025		https://mp.	MoE	Reasoning	"Based on Turbo S, by introducing technologies such as long thinking chains, retrieval er
Hunyuan Turbo S	Tencent	https://cloud.ten	*389*	*7000*	18:1	5.5	89.5	79	57.5	synthetic, web-scale	Feb/2025		https://mp.	MoE		Fast thinking ("Instant reply"). "This is also the first time in the industry that the Mamba
Phi-4-multimodal	Microsoft	https://huggingfa	5.6	6100	1,090:1	0.6				synthetic, web-scale	Feb/2025		https://hug	Dense		"Training data: 5T tokens, 2.3M speech hours, and 1.1T image-text tokens" Announce: h
Phi-4-mini	Microsoft	https://huggingfa	3.8	5000	1,316:1	0.5	67.3	52.8	30.4	synthetic, web-scale	Feb/2025		https://hug	Dense		"Phi-4-mini's training data includes a wide variety of sources, totaling 5 trillion tokens, a
Mercury Coder Sma	Inception	https://chat.ince	*40*	*5000*	125:1	1.5				synthetic, web-scale	Feb/2025		https://ww	*Dense*		Diffusion large language model (dLLM). Very low 'IQ' performance (0/5 on all ALPrompts
QwQ-Max-Preview	Alibaba	https://chat.qwer	325	20000	62:1	8.5				synthetic, web-scale	Feb/2025		https://qwe	Dense	Reasoning	"As a sneak peek into our upcoming QwQ-Max release, this version offers a glimpse of it
Claude 3.7 Sonnet	Anthropic	https://claude.ai/	*175*	*20000*	115:1	6.2			84.8	synthetic, web-scale	Feb/2025		https://asse	Dense	Reasoning SOTA	Knowledge cutoff now November 2024 (was April 2024). "the first hybrid reasoning mod
Moonlight	Moonshot AI	https://huggingfa	16	5700	357:1	1.0	70	42.4		synthetic, web-scale	Feb/2025		https://gith	MoE		"Scaling law experiments indicate that Muon achieves ~ 2× computational efficiency co
S2	Figure		7	*2000*	286:1	0.4				synthetic, web-scale	Feb/2025		https://ww	Dense		Likely based on OpenVLA 7B (Jun/2024, based on Llama 2 7B) or Molmo 7B-O (Sep/2024
S1	Figure		0.08	*1*	13:1	0.0				special	Feb/2025		https://ww	Dense		"high quality, multi-robot, multi-operator dataset of diverse teleoperated behaviors, ~50
Baichuan-M1-14B	Baichuan	https://huggingfa	14	20000	1,429:1	1.8				synthetic, web-scale	Feb/2025		https://arxiv	Dense		Medical LLM. Huge increase to 20T tokens for 14B params standard.
Evo 2	Arc Institute	https://github.co	40	8800	220:1	2.0				special	Feb/2025		https://gith	Dense		"Evo 2 is a state of the art DNA language model for long context modeling and design. Ev
R1 1776	Perplexity	https://huggingfa	*685*	*14800*	22:1	10.6	90.5			synthetic, web-scale	Feb/2025		https://ww	MoE	Reasoning	Censorship reduced, based on DeepSeek-R1.
Grok-3	xAI	https://grok.com/	*928*	*36200*	40:1	19.3		79.9	84.6	synthetic, web-scale	Feb/2025		https://x.ai/	*MoE*	Reasoning SOTA	https://x.ai/blog/grok-3 My full analysis: https://lifearchitect.ai/whats-in-grok/
Mistral Saba	Mistral	https://console.m	24	*8000*	334:1	1.5	*81*			web-scale	Feb/2025		https://mist	Dense		"Mistral Saba is a 24B parameter model trained on meticulously curated datasets from a
Salamandra	Barcelona Superco	https://github.co	40	9000	225:1	2.0				web-scale	Feb/2025		https://arxiv	Dense		"The final [pre-training] dataset is composed of 55.51% FineWeb-Edu, 25.32% Colossal C

Quelle: lifearchitect.ai/models - ***März 2025***

Grundprinzip GPT



Länge des Eingabetextes ist begrenzt!



Durch GPT findet eine probabilistische, kontextbezogene **Wortvorhersage** statt.

Maßnahmen zur Anpassung

Es existieren verschiedene Techniken, um die Funktionalitäten des LLMs für die eigene Anwendung anzufertigen.

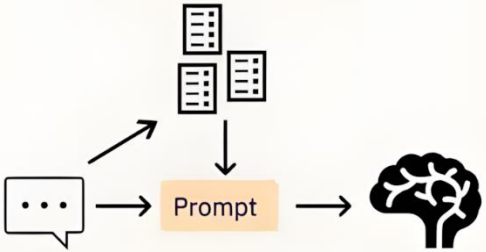
Prompt Engineering

Prompt

- *Standardanweisung*
- *Spezielle Anweisungen*
- *Beispiele:*
  - *Beispiel 1 – Antwort 1*
  - *Beispiel 2 – Antwort 2*
  - *...*

Füge spezielle Anweisungen oder Beispiele dem Prompt hinzu

Retrieval Augmented Generation (RAG)



Stimme Input mit eigenen Daten ab und nutze passende Daten als neuen Promptkontext

Fine-Tuning

Nutze eigene Daten um ausgewählte ParameterIm LLM neu zu trainieren

Grad der Komplexität / Rechnerischer Aufwand

https://datasolut.com/was-ist-ein-large-language-model

Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ✗

Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✓

Quelle: Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, Wei et al., Jan. 2023

RAG Beispiel Angebotsvorbereitung

```
Du bist Mirko Schlaumeier, Kundenberater für den Bereich Blankstahl, Qualitätsstahl und hast eine Email-Anfrage erhalten.

Start der Emailanfrage des potenziellen Kunden:
Guten Tag Herr Schlaumeier,

wir hatten schon über meine Bestellung telefonisch gesprochen.
Bitte machen Sie mir doch eine Angebot fertig mit den besprochenen Positionen.

Vielen Dank

Hans Mustermann

Ende der Email
```

```
Eine Suche in der Kommunikation mit dem Kunden hat folgende Inhalte gefunden:

- Anruf 15.02.2025 14:24 Uhr von Herrn Mustermann
- er war sehr zufrieden und hat sich über die Sonderangebote bzgl. Kurzlängen informiert
- es kommen 500 Stück zu 3m Länge in Frage
- Art. Nr. 705171

Ende der Kommunikationsdaten
```

RAG

```
Eine Suche in der Artikeldatenbank hat folgendes Ergebnis zu den Artikelnummern ergeben:

Artikelnummer 705171
Bezeichnung: 25CrMo4+QT+SH h9 beids. gefast
Preis: 1100 € / Tonne
Aktueller Lagerbestand: 574
Form: Rd 12

Ende des Suchergebnisses zu Artikelstammdaten

Prüfe, ob die geforderte Menge an Produkten verfügbar ist.
Sofern die geforderte Menge verfügbar ist, schreibe einen kurzen Angebotstext,
der das Produkt beschreibt. Nutze eine geeignete Anrede und Schlußformel.
```

RAG

Assistant

Guten Tag Herr Mustermann,

vielen Dank für Ihren Anruf und Ihr Interesse an unseren Blankstahl-Produkten. Ich freue mich, Ihnen das folgende Angebot unterbreiten zu können.

Sie interessieren sich für 500 Stück zu je 3 Metern Länge des Artikels 705171, Bezeichnung: 25CrMo4+QT+SH h9 beidseitig gefast. Dieses Produkt bieten wir aktuell zum Preis von 1100 € pro Tonne an. Der aktuelle Lagerbestand beträgt 574 Einheiten, sodass die von Ihnen gewünschte Menge von 500 Stück erfüllt werden kann.

Sollten Sie weitere Fragen haben oder eine Bestellung aufgeben wollen, stehe ich Ihnen selbstverständlich jederzeit zur Verfügung.

Ich freue mich darauf, bald wieder von Ihnen zu hören.

Mit freundlichen Grüßen,

Mirko Schlaumeier
Kundenberater Blank- und Qualitätsstahl

how many Rs are in strawberry

OI **deepseek-r1:8b**

<think>
Okay, so I'm trying to figure out how many times the letter "R" appears in the word "strav
down the word and examining each letter one by one.

The word is S-T-R-A-W-B-E-R-R-Y. Hmm, I should probably write it out in a more linear w
breaking it down:

1. S
2. T
3. R
4. A
5. W
6. B
7. E
8. R
9. R
10. Y

Wait, let me make sure I have the letters in order. Sometimes when I'm focusing on parts
a letter. Let me write it again: S, T, R, A, W, B, E, R, R, Y.

I think I'm confident with this count now. It's always good to double-check because some
words are spelled differently or if I'm tired. But in this case, I'm pretty sure it's three R's.
</think>

The letter "R" appears three times in the word "strawberry."

Step-by-Step Explanation:

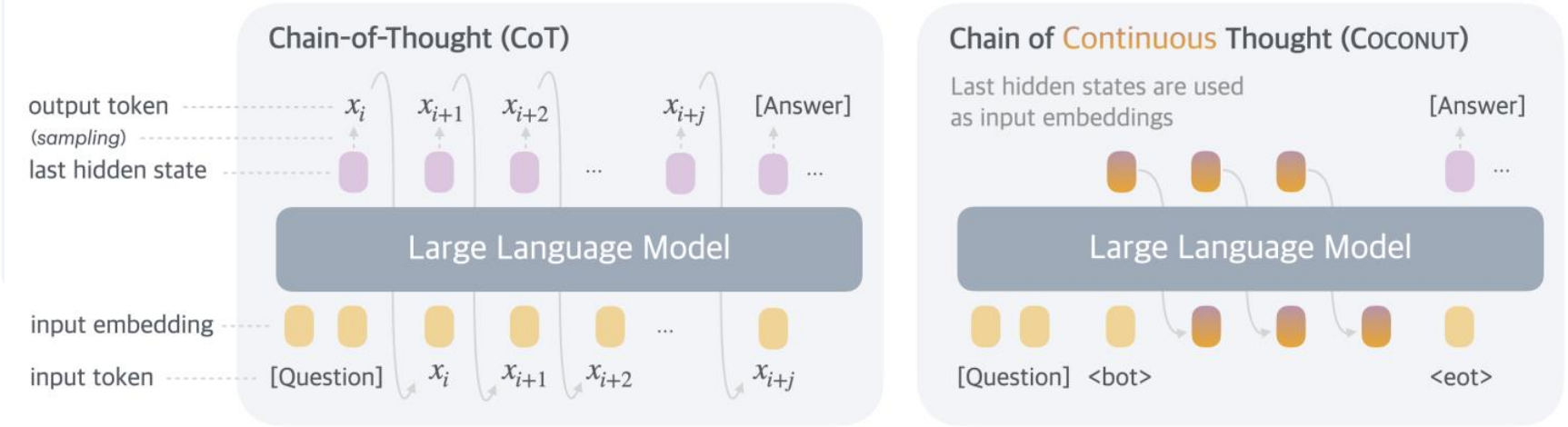
1. **Write Down the Word:** Start by writing out the word to visualize each letter clearly.
   - S T R A W B E R R Y
2. **Identify Each Letter Position:**
   - 1: S
   - 2: T
   - 3: R
   - 4: A
   - 5: W
   - 6: B
   - 7: E
   - 8: R
   - 9: R
   - 10: Y
3. **Count the "R"s:**
   - Position 3: R (1st occurrence)
   - Position 8: R (2nd occurrence)
   - Position 9: R (3rd occurrence)
4. **Total Count:** There are three "R" letters in the word.

Answer: The letter "R" appears three times in "strawberry."

Training Large Language Models to Reason in a Continuous Latent Space

Shibo Hao[1,2,*], Sainbayar Sukhbaatar[1], DiJia Su[1], Xian Li[1], Zhiting Hu[2], Jason Weston[1], Yuandong Tian[1]
[1]FAIR at Meta, [2]UC San Diego
*Work done at Meta

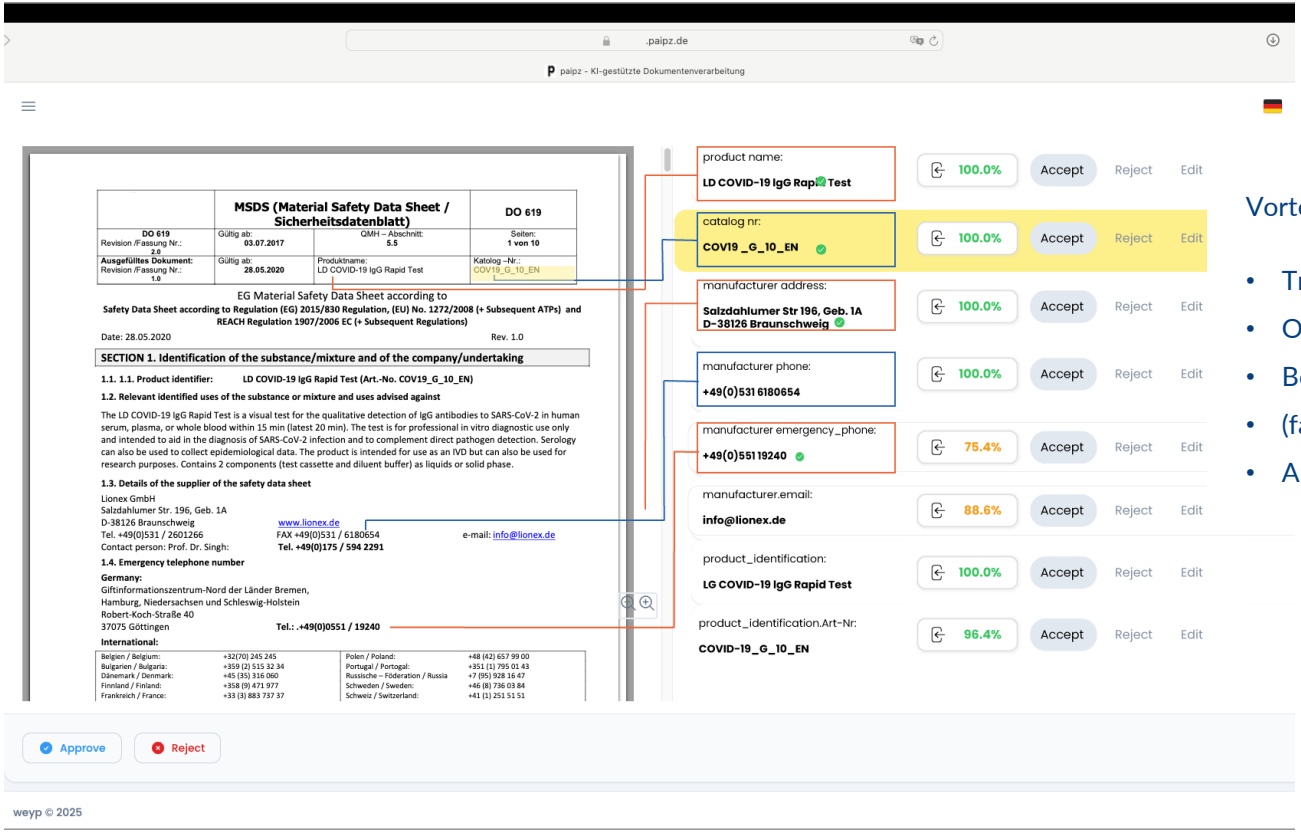
Large language models (LLMs) are restricted to reason in the "language space", where they typically express the reasoning process with a chain-of-thought (CoT) to solve a complex reasoning problem



Chain-of-Thoughts auf Basis der **internen** Repräsentation des Ergebnisses.
Es werden **keine Wörter** für die Zwischenergebnisse generiert.

Quelle: Training Large Language Models to Reason in a Continuous Latent Space, Hao et al., Dec 2024

Beispiel Datenextraktion



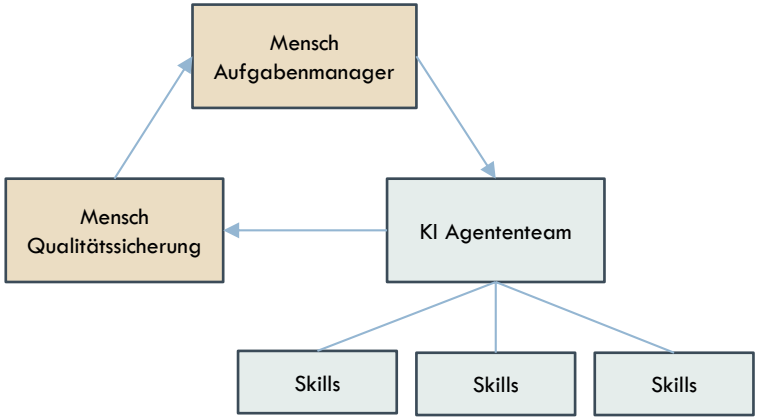
Quelle: paipz.de

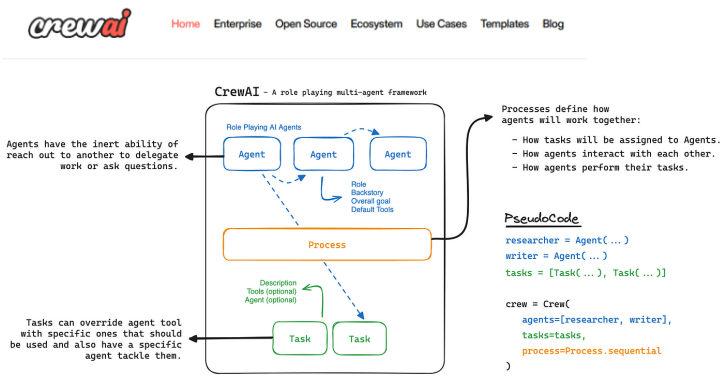
Vorteile KI-basierter Extraktion:

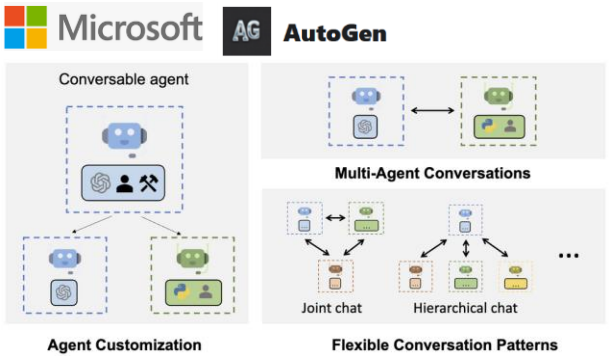
- Training auf bestehenden PDF-Sammlungen
- Ortsinvariant, d.h. kein Oldschool-Templating
- Berücksichtigung der Fachterminologie
- (fast) sprachunabhängig
- Automatisierungsgrad individuell einstellbar

Arbeitsteilung KI

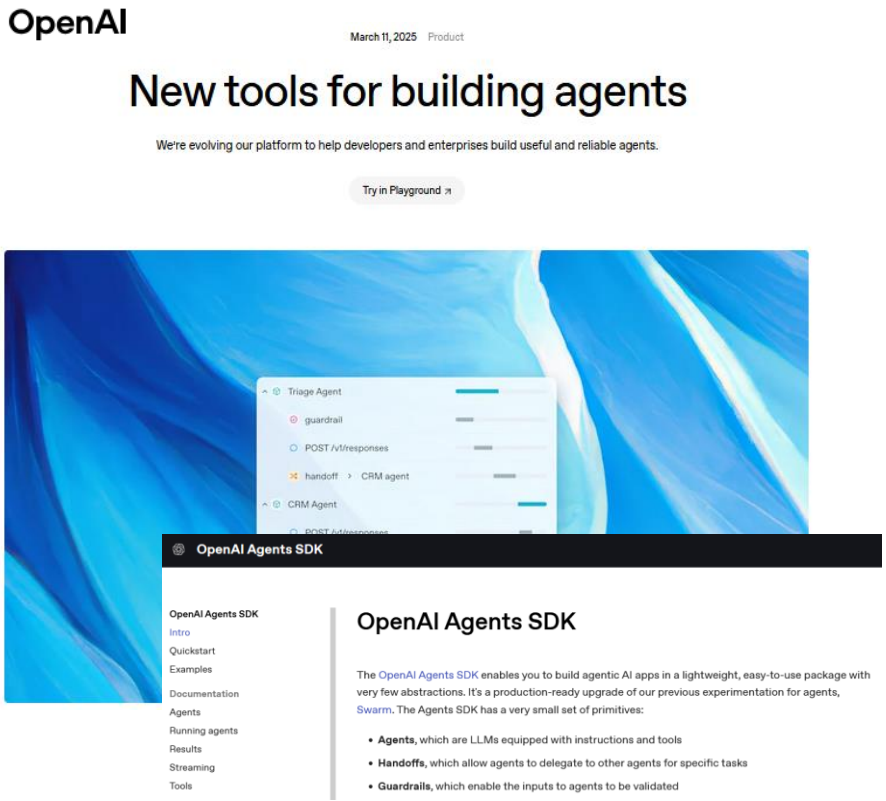
- Multi-Agenten-Systeme
- Arbeitsteilung/Kollaboration
- Verwendung von unterschiedlichen LLMs/SLMs
- Open-Source, Fully-Local
- Cloud-basierte Skalierung von Services







Quelle: techcommunity.microsoft.com/t5/ai-azure-ai-services-blog/building-a-multimodal-multi-agent-framework-with-azure-openai/ba-p/4084007



SDK ist Open-Source!