Thomas Jörg, LFT KI Baden-Württemberg, Informatiklehrer am Kepler Gymnasium Pforzheim

- Lokale SLM (LLM) Modelle ...

Gemma AI Models | Phi-3 | Aya 23 cohere | MISTRAL AI_ | Meta LLAMA 3

dfki ai | hessian.AI | TECHNISCHE UNIVERSITÄT DARMSTADT | COMMON CRAWL | EleutherAI | DISCO RESEARCH LAB

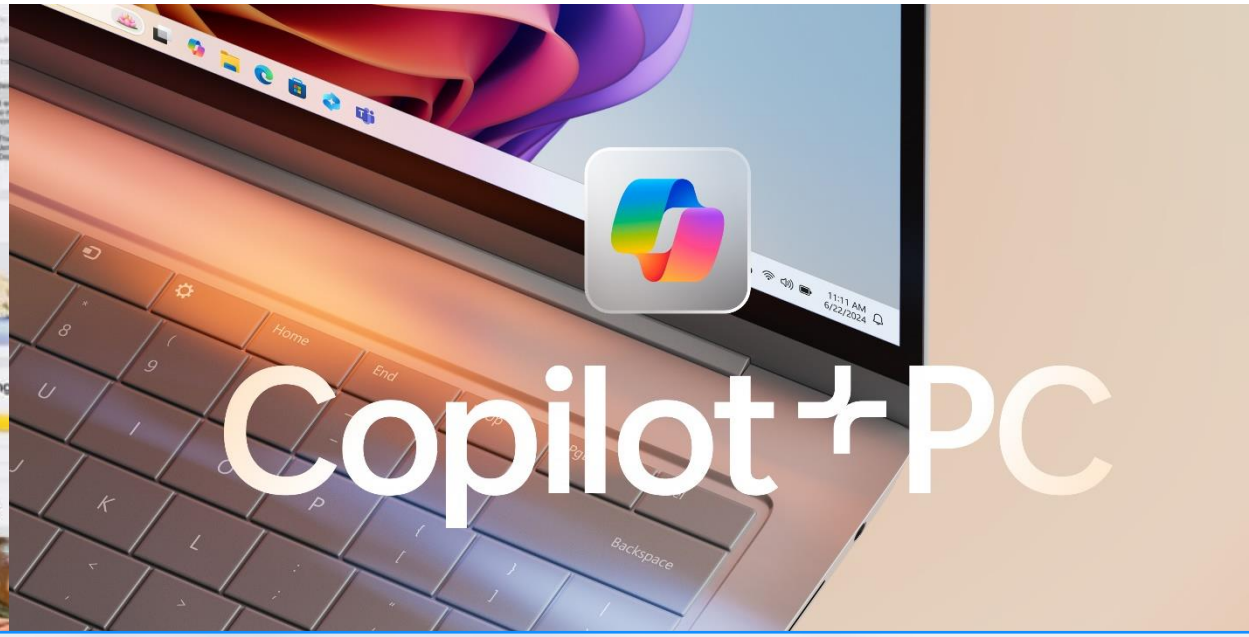
- ... versus ‚Big Players‘:

Lokale SLM (LLM) Modelle...

- Use cases: Wozu sind diese SLM / LLM gut?
- Welche (Datei-)Eigenschaften haben diese Modelle?
- Welche Hardware braucht man?
- Welche Software braucht man?
- Wo liegen (aktuell) die Grenzen? Wo geht die Entwicklung hin?

Warum lokale Sprachmodelle, die weniger leistungsfähig sind?

- Bei den ‚Big Three' mehr als problematisch: Datenschutz / Privacy
- Bei den ‚Big Three' immer eine Internetverbindung nötig.
- Für viele Aufgaben ist das umfassende Wissen der ‚Großen' nicht nötig



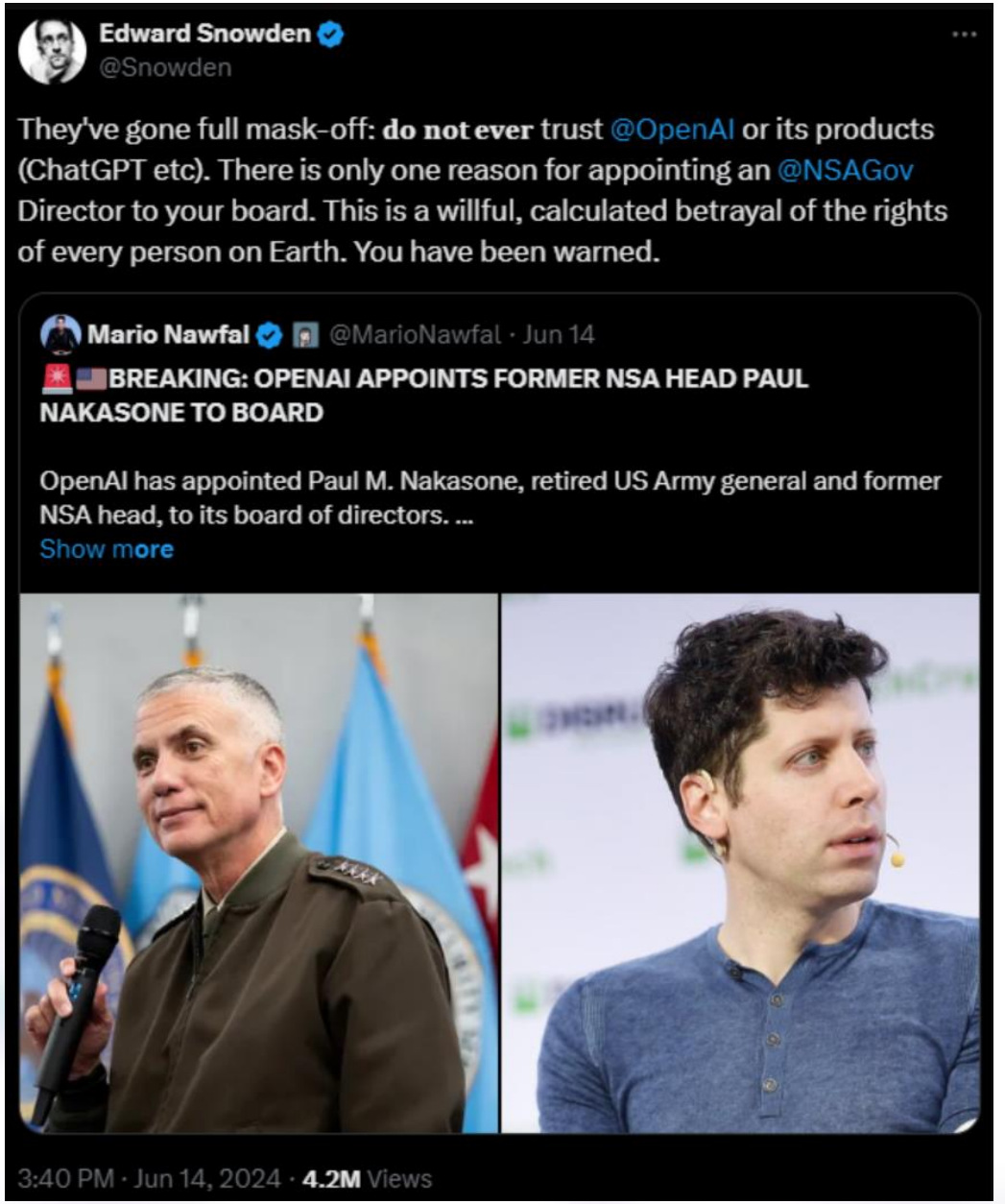
- *Stand 21. Juni 2024: zumeist noch ‚Zukunftsmusik' …*

Warum lokale Sprachmodelle,?

„He concurrently served as the director of the National Security Agency NSA"

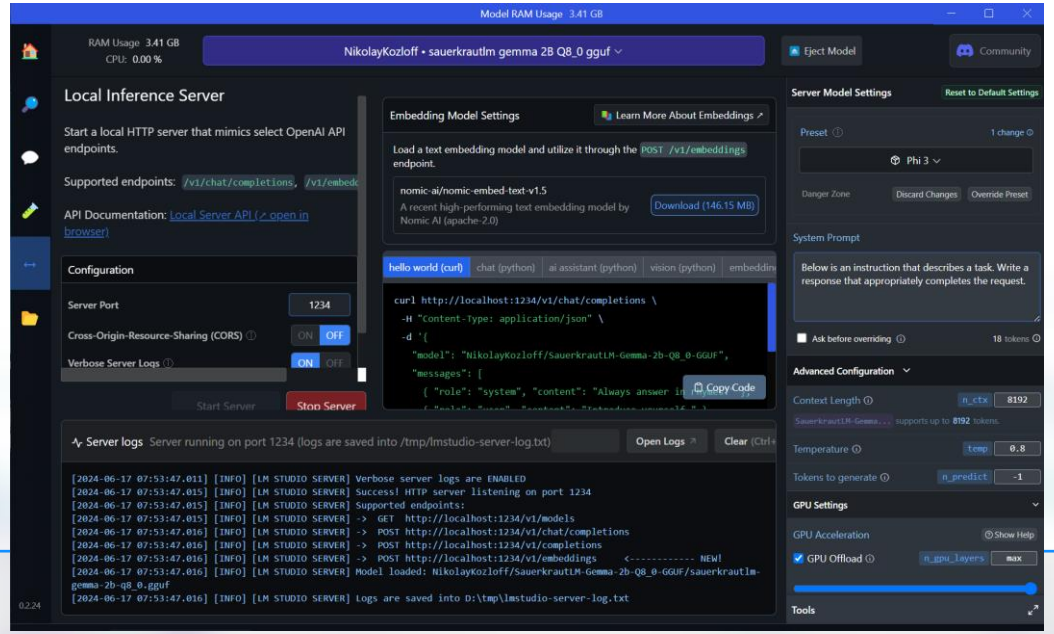
„Nach Beendigung seiner militärischen Laufbahn im Februar 2024 übernahm Nakasone im Juni 2024 einen Posten im Verwaltungsrat der auf KI spezialisierten Firma OpenAI."

https://the-decoder.de/edward-snowden-haelt-chatgpt-und-openai-nach-nsa-verbindung-fuer-nicht-mehr-vertrauenswuerdig/

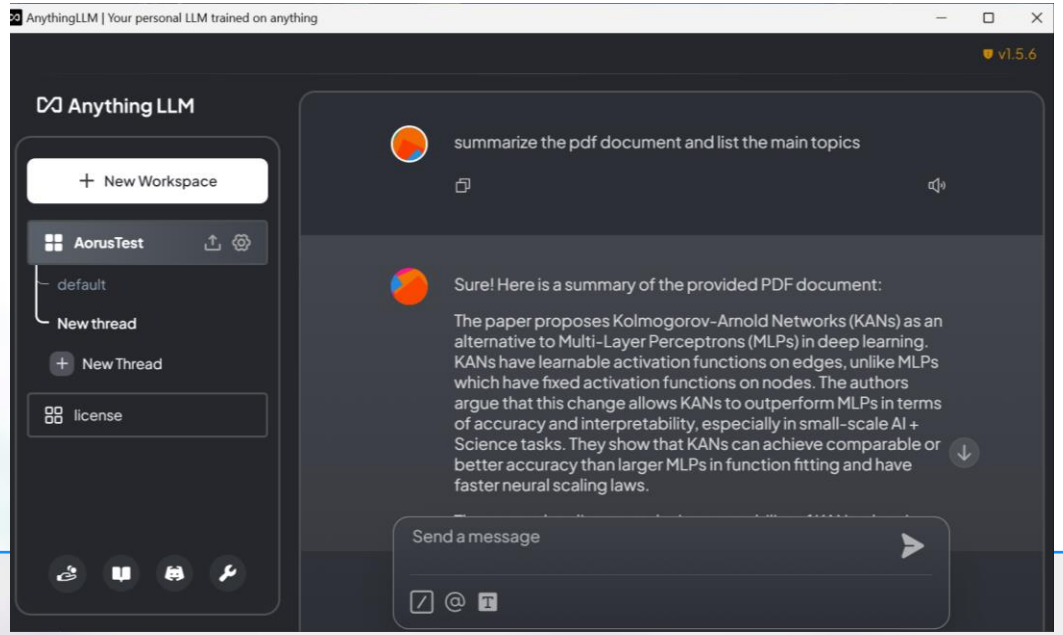


Was funktioniert ‚jetzt'? (21. Juni 2024)

- Opensource-LLMs (werden auch von Apple/Microsoft verwendet)
- Software, um LLMs/SLMs zu hosten
- Software für RAG (‚Retrieval Augmentation Generation')



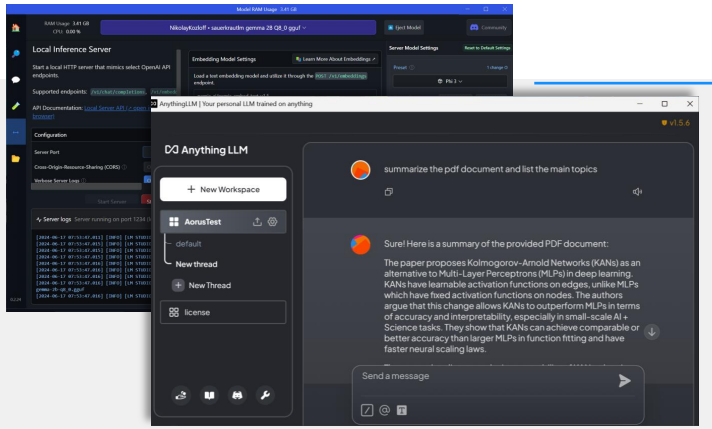
LM Studio



AnythingLLM

Was funktioniert ‚jetzt'? (21. Juni 2024)

DEMO LM Studio & Anything LLM.



- LM Studio (SLM Hosting):

  https://lmstudio.ai/

- Anything LM (Business Intelligence):

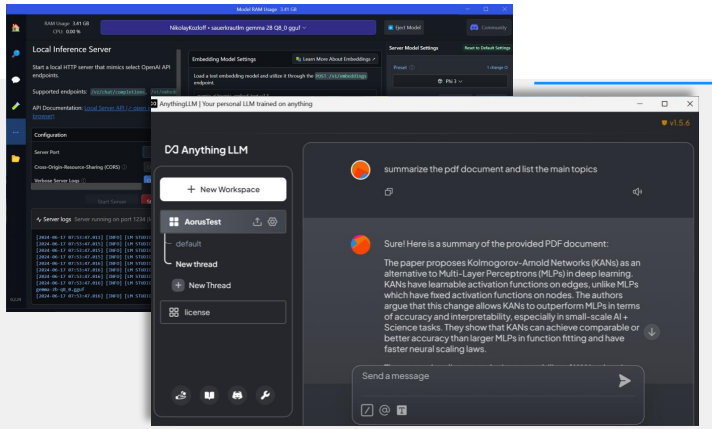
  https://useanything.com/

- Huggingface: 🤗 "riesiger Basar fuer KI"

  https://huggingface.co/models?pipeline_tag=text-generation&sort=downloads

Was funktioniert ‚jetzt'? (21. Juni 2024)

DEMO LM Studio & Anything LLM.



- **LM Studio:**

  Suche, Herunterladen, Grafikkarten-Offload, Server,

  *Achtung: **Context length**!*

sollten gleich sein.

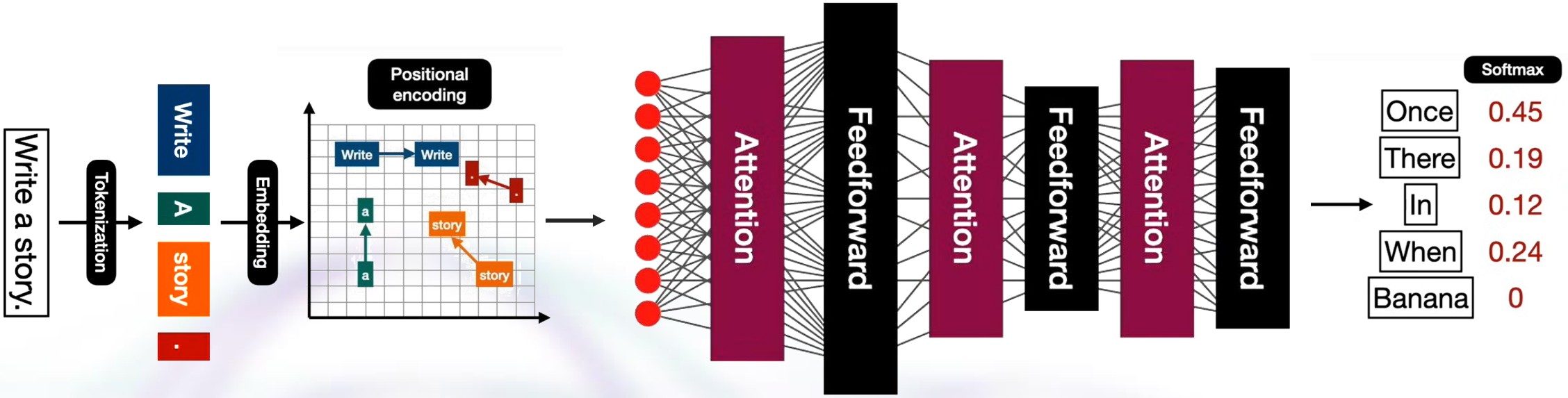
- **Anything LM:**

  *General Setup (**Token Context Window**!), Workspace, Document Database,*

  *Embeddings, Whisper, **Agent Configuration**!*

Was ist ein LLM (Large Language Model)?

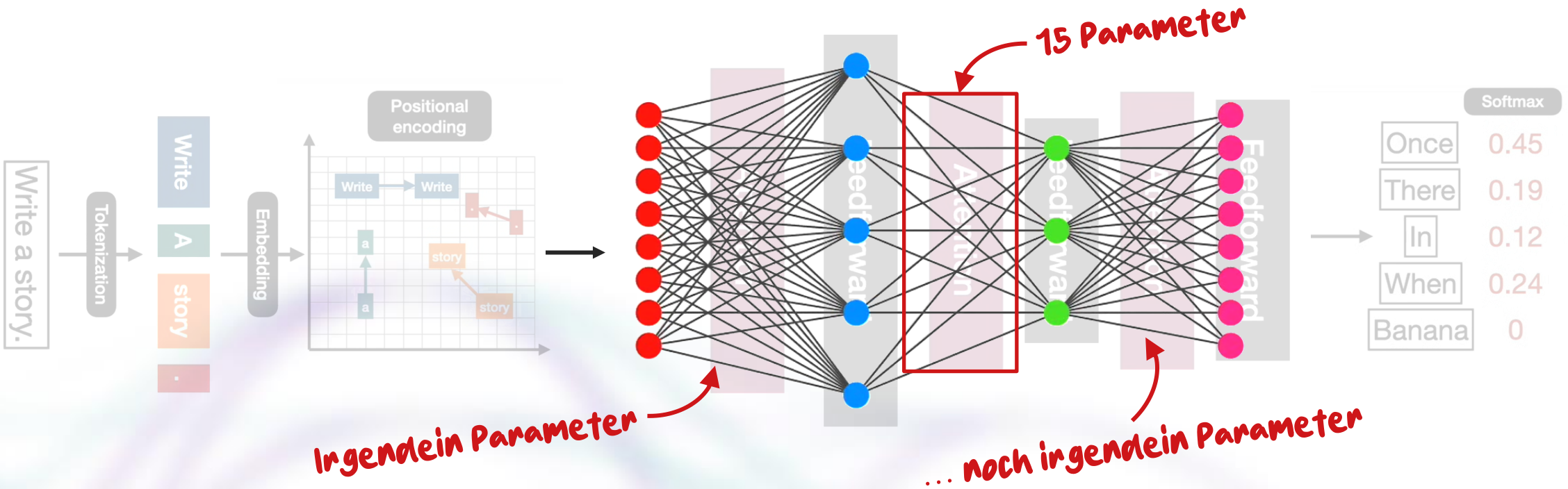
- Ein neuronales Netz mit spezieller Architektur (Attention-Mechanismus)



What are Transformer Models and how do they work? (Luis Serrano)
https://www.youtube.com/watch?v=qaWMOYf4ri8

Was ist ein LLM (Large Language Model)?

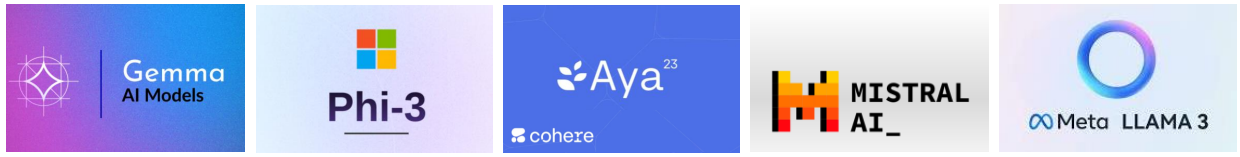
- Ein neuronales Netz mit spezieller Architektur (Attention-Mechanismus)



What are Transformer Models and how do they work? (Luis Serrano)
https://www.youtube.com/watch?v=qaWMOYf4ri8

Welche Basis-Modelle sind (21. Juni 2024) besonders nützlich?

FACHBEGRIFF!



- *Tiny 2B, 3B:* Phi-3 mini
- *Small-Medium 7B, 8B:* LlamaV3 8B, Mistral 7B, Aya23 8B
- *Medium: 14B, 35B:* Phi-3 medium (Aya23 35B)

- *NICHT betrachtet werden die Big-Models: 70B/140B usw.*

  aufgrund der Dateigröße, deshalb auf ‚normalen' Rechnern nicht lauffähig.

Welche Finetuned-Modelle betrachten wir in diesem Vortrag?

WICHTIG!

- Nur sogenannte ‚Instruct'-Modelle, weil auf Chat/Befehle hin trainiert.
- Nur solche Modelle, welche auf deutsche Sprache trainiert wurden.

Welche Modelle sind das ganz konkret (Achtung! Empfehlung mit Noten 😉):

- Phi-3-mini-4k-instruct [Note 2-]
- Llama3-DiscoLeo-Instruct-8B [1-] | occiglot-7b-eu5-instruct [2] | aya23-8B [2+]
- Phi-3-medium-128k-instruct [1] | aya23-35B [2- weil eigentlich zu groß]

Welche Hardware-Vorrausetzungen gibt es? *Teil 1 von 2*

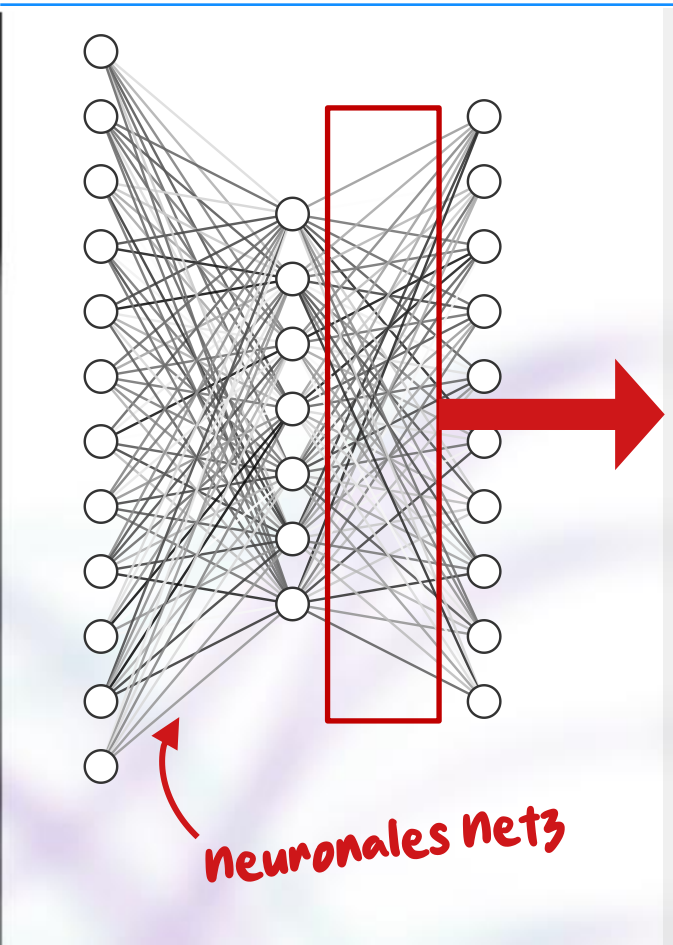
- „Das natürliche Biotop von neuronalen Netzen sind GPU und NPU/TPU“

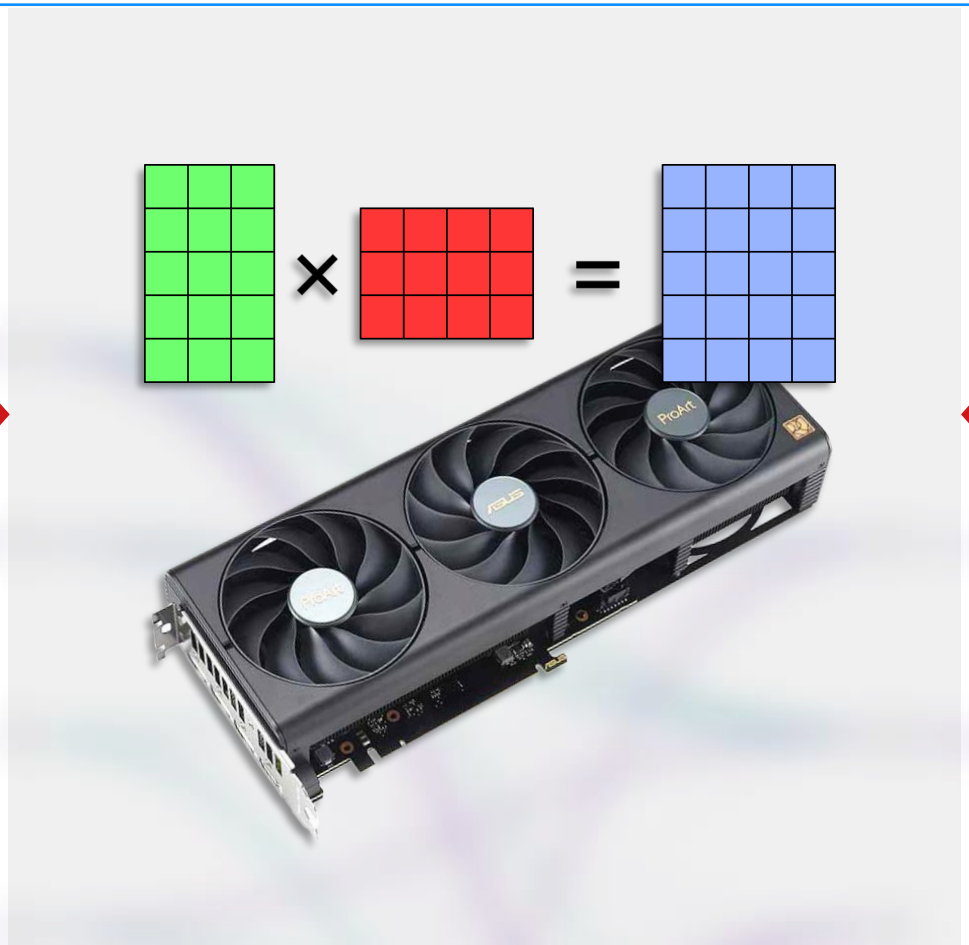
- Glücklicherweise sind NN. auf NVIDIA-Gaming-GraKas lauffähig (GPU)

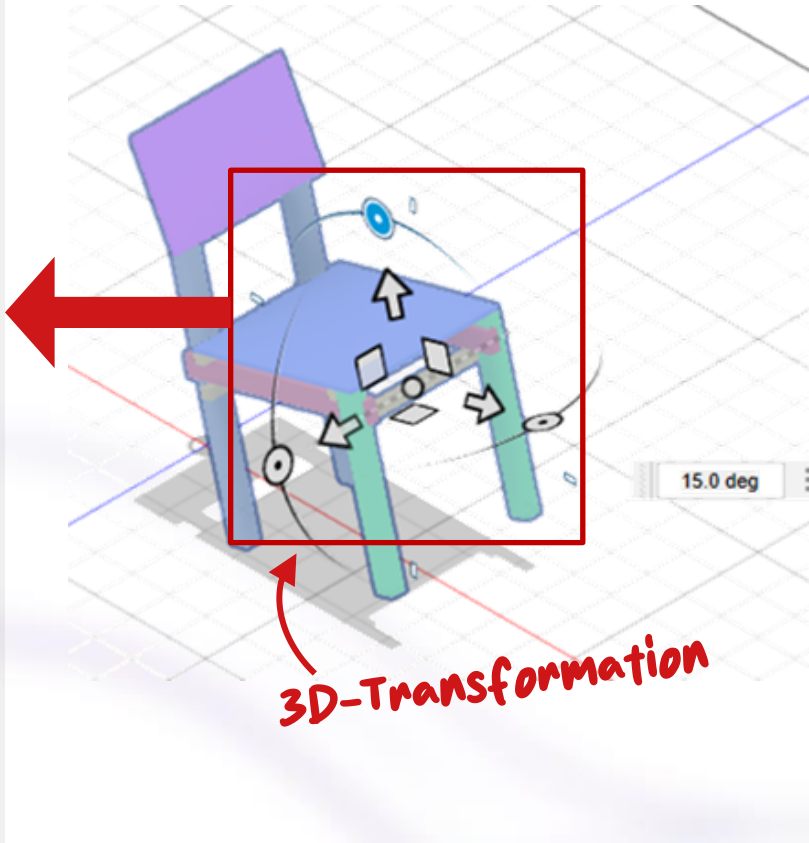
- *Stand 21.6.2024: Hauptgewicht auf NVIDIA, Nebengewicht auf Apple M2 / M3*
- *Hoffnungsträger: Snapdragon mit integrierter NPU (... to be continued ...)*

Welche Hardware-Vorrausetzungen gibt es? *Teil 1 von 2*

- Warum GPU? ***Beide Bereiche nutzen parallelisierte Matrizenrechnungen***







Welche Hardware-Vorrausetzungen gibt es? *Teil 2 von 2*

- VRAM (bei NVIDIA) und ‚Shared RAM' bei Apple / Microsoft
- Je mehr desto besser. Aber mindestens (!!) **8GB** dediziert für NPU / GPU.
- Obergrenze derzeit bei NVIDIA: 24 GB Desktop (4090), 16 GB Laptop (3080 / 3080Ti)
- Obergrenze derzeit bei Apple: Geldbeutel, bzw. 128GB (min. 5724 €)
- Obergrenze derzeit bei Snapdragon: 16 GB („erste Welle Copilot+")

Welche Quantisierungen sind für ein lokales Modell vernünftig?

- Q8_0 | Q6_K | Q5_K_M 😎 👌

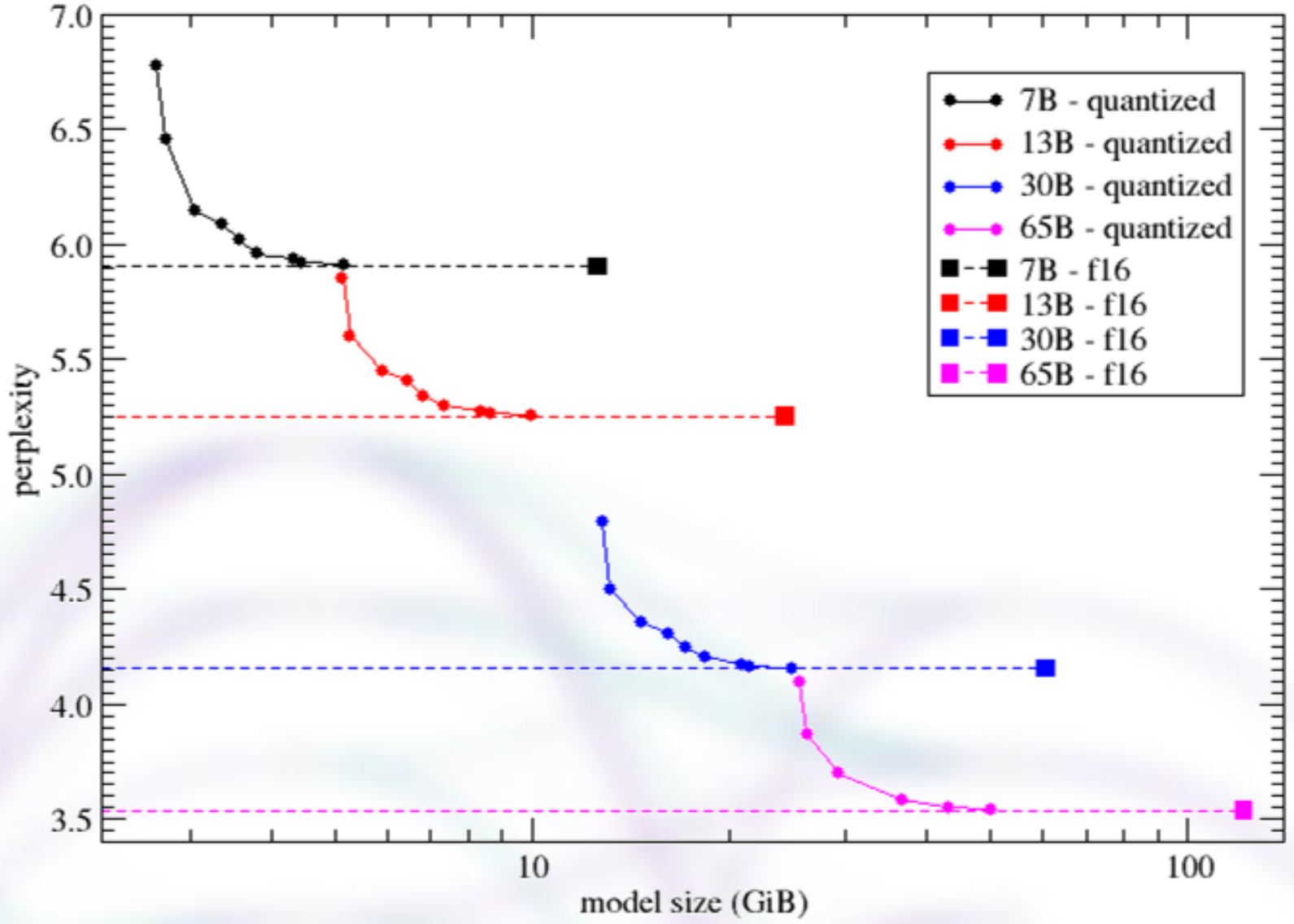
- https://huggingface.co/dranger003/c4ai-command-r-plus-iMat.GGUF
  https://github.com/ggerganov/llama.cpp#quantization
  https://github.com/ggerganov/llama.cpp/pull/1684#issue-1739619305

Wie bestimmt man das?

- Perplexity PPL (= Modellunsicherheit): Qualitätsmaß für LLMs
- Je niedriger die Perplexity, desto genauer/sicherer das Modell
- Beispiel: *„Der Himmel ist …“* → Niedrige PPL: *„blau“* | Hohe PPL: *„lila“*
- Satz von Standard-“Ratefragen“, um Modelle zu vergleichen

Welche Quantisierungen sind für ein lokales Modell vernünftig?

2-6 Bit Quantisierung



Welche Quantisierungen sind für ein lokales Modell vernünftig?

Measure	Q2_K	Q3_K_S	Q3_K_M	Q4_K_S	Q4_K_M	Q5_K_S	Q5_K_M	Q6_K	FP16
Perplexity	6.77	6.46	6.15	6.02	5.96	5.94	5.92	5.91	5.91
file size (GB)	2.67	2.75	3.06	3.56	3.80	4.33	4.45	5.15	13.0
ms/token, RTX-4080	15.5	18.6	17.0	15.5	16.0	16.7	16.9	18.3	60

- Q6_K oder Q5_K_M haben nur sehr geringe Verluste,
- Geschwindigkeit akzeptabel,
- Dateigröße im Speicher mit max. 16GB VRam realistisch.

Beispiel Q5_K_M-Quantisierung:

EXAKTER WERT

Quantisierung:

$$\begin{pmatrix}0.10\\0.25\\0.50\\0.75\\0.90\end{pmatrix} \xrightarrow{5\text{-}Bit\ Quantisierung,\ 2^5 = 32\ Stufen\ (von\ 0\ bis\ 31)} \begin{pmatrix}0.10\times 31\\0.25\times 31\\0.50\times 31\\0.75\times 31\\0.90\times 31\end{pmatrix} = \begin{pmatrix}3.10\\7.75\\15.50\\23.25\\27.9\end{pmatrix} \xrightarrow{Runden} \begin{pmatrix}3\\8\\16\\23\\28\end{pmatrix}$$

QUANTISIERTER WERT

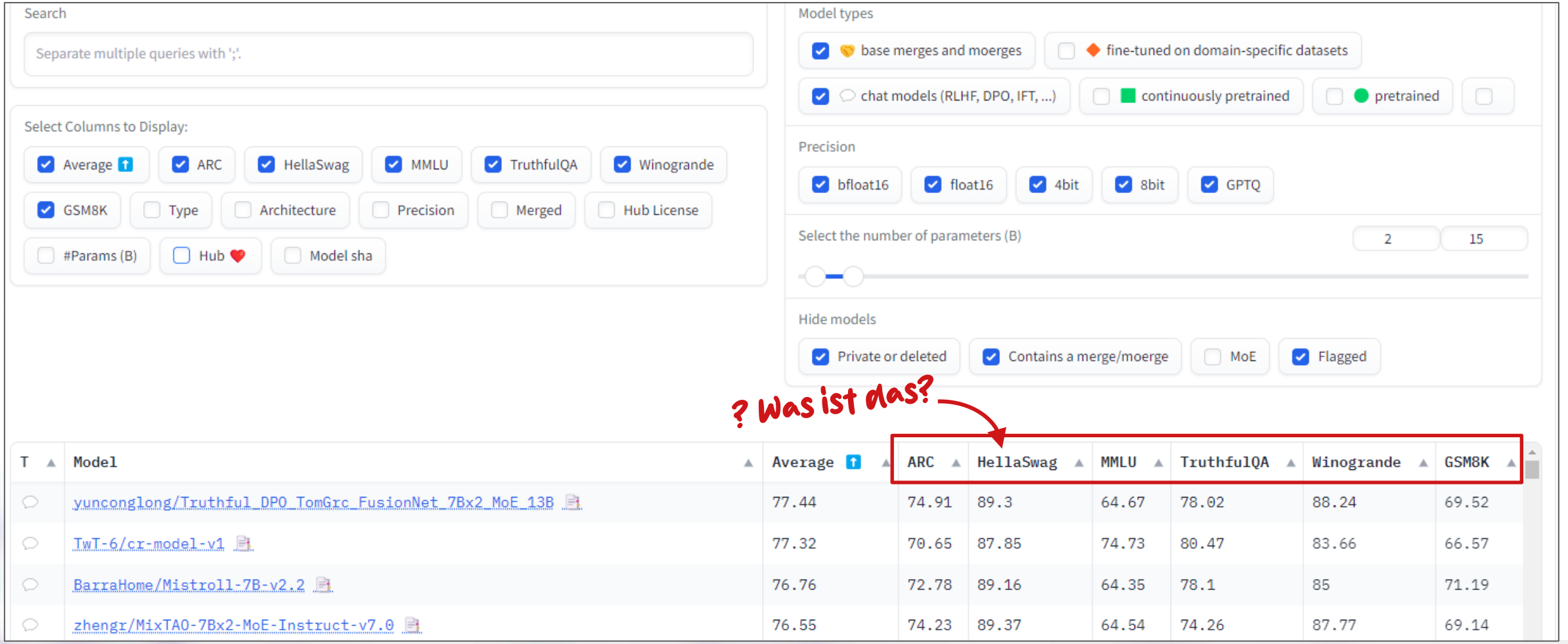
Vorher zu nachher

Zurückrechnen:

$$\begin{pmatrix}3\\8\\16\\23\\28\end{pmatrix} \xrightarrow{Zurückrechnen\ in\ Floating\ Point} \begin{pmatrix}3/31\\8/31\\16/31\\23/31\\28/31\end{pmatrix} = \begin{pmatrix}0.097\\0.258\\0.516\\0.742\\0.903\end{pmatrix}$$

DEKOMPRIMIERTER WERT

Gibt es eine ‚Bestenliste' der aktuellen Sprachmodelle? Ja: Das „Leaderboard"



T	Model	Average	ARC	HellaSwag	MMLU	TruthfulQA	Winogrande	GSM8K
	yunconglong/Truthful_DPO_TomGrc_FusionNet_7Bx2_MoE_13B	77.44	74.91	89.3	64.67	78.02	88.24	69.52
	TwT-6/cr-model-v1	77.32	70.65	87.85	74.73	80.47	83.66	66.57
	BarraHome/Mistroll-7B-v2.2	76.76	72.78	89.16	64.35	78.1	85	71.19
	zhengr/MixTAO-7Bx2-MoE-Instruct-v7.0	76.55	74.23	89.37	64.54	74.26	87.77	69.14

- https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard

LLM Abschluss – Pruefung

- **Winogrande: Bewertet die Fähigkeit von LLMs, Nuancen in der Sprache zu verstehen und Mehrdeutigkeiten aufzulösen.**

  Beispiel: „*Der Anwalt überzeugte den Richter mit seinem Fall*“.

  Wähle mögliche Bedeutungen von "Fall":

  (a) ein juristischer Fall (b) ein Sturz

- **GSM8k (Grade School Math 8K): bewertet die Fähigkeit von LLMs, mathematische Textaufgaben zu lösen, die für Grundschüler konzipiert sind.**

  Beispiel: *„Ein Bauer hat 12 Kühe und jede Kuh produziert 10 Liter Milch pro Tag.“*

  Wie viel Milch produzieren alle Kühe zusammen an einem Tag?

  Lösung: 12 Kühe * 10 Liter/Kuh = 120 Liter

LLM Abschluss – Pruefung

- **HellaSwag: präsentiert dem Modell unvollständige Sätze und fordert es auf, das beste Ende auszuwählen. Für Menschen oftmals einfach, aber für Maschinen schwierig.**

Beispiel: *„Ein Kind steht vor einem Eiswagen und hält Geld in der Hand. Es… “*

Finde die beste Endung:

(a) kauft ein Eis. **(b) geht zur Schule.** **(c) liest ein Buch.**

- **TruthfulQA: misst, wie gut ein LLM wahre und falsche Aussagen erkennen kann. Es präsentiert teils Fehlinformationen und bewertet, wie gut das Modell die Wahrheit erkennt.**

Beispiel: *„Ist es wahr, dass Zitronen mehr Zucker enthalten als Erdbeeren?“*

(a) Ja **(b) Nein**

LLM Abschluss – Pruefung

- MMLU (Massive Multitask Language Understanding): bewertet eine Vielzahl von Aufgaben in verschiedenen Bereichen zu verstehen, z.B. aus Mathe, Geschichte, Informatik und Jura.

Beispiel: *„Welches Ereignis markierte den Beginn des Zweiten Weltkriegs in Europa?“*

(a) Der Versailler Vertrag
(b) Der Einmarsch Deutschlands in Polen
(c) Angriff auf Pearl Harbor
(d) Der Angriff auf die Sowjetunion

- ARC (AI2 Reasoning Challenge): testet die Fähigkeit eines LLM, wissenschaftliches Konzeptverständnis zu zeigen, wie es für Schüler der 3. bis 9. Klasse typisch ist.

Beispiel: *„Welcher der folgenden Faktoren trägt am meisten zur Entstehung von Wind bei?“*

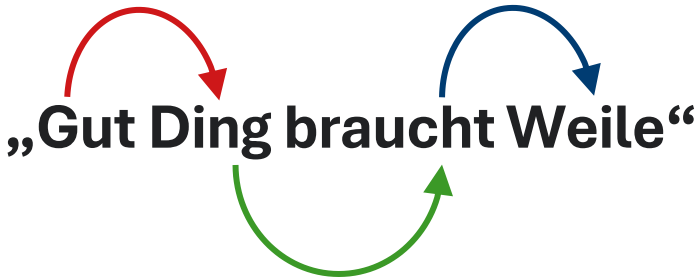
(a) Erdrotation
(b) Lufttemperatur-Unterschiede
(c) Anziehungskraft des Mondes
(d) Wolkenmenge am Himmel

Use cases: Wozu sind diese SLM / LLM gut?

Gemma AI Models | Phi-3 | Aya 23 cohere | MISTRAL AI_ | Meta LLAMA 3

- Chat / Diskussion (klassischer Chatbot)
- Übersetzungen Deutsch – Französisch – Englisch – Italienisch …
- RAG: PDFs/Ebooks/Mails/Tabellen zusammenfassen / untersuchen / übersetzen
- RAG: Texte von eingebetteten Webseiten zusammenfassen
- Coding Assistant: Programmier-Unterstützung
- *Audiofiles transkribieren (gehobener Beta-Status)*
- *Agents: Webscraping, Dokumentenanalyse, Dateispeicherung (noch ziemlich „Beta“)*

3. Attention | Wortbezüge: Alle Wörter werden mit allen kombiniert.

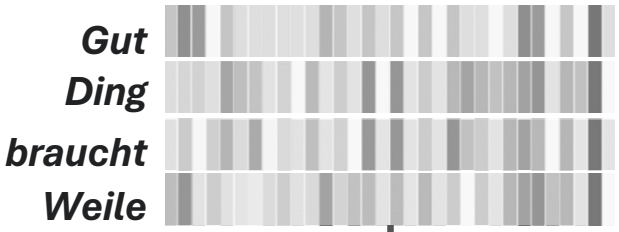


- Senkrecht, als Spaltenbezeichnung: QUERY-Wörter, **für** die die Scores* bestimmt werden
- Waagerecht, als Zeilenbezeichnung: Key-Wörter, **mit** denen die Scores* bestimmt werden.

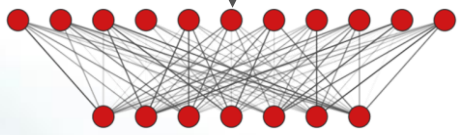
KEY →
QUERY ↓

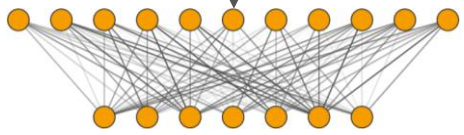
	Gut	*Ding*	*braucht*	*Weile*
Gut	0.3	**0.5**	0.1	0.1
Ding	0.2	0.3	**0.4**	0.1
braucht	0.2	0.1	0.2	**0.5**
Weile	0.1	0.3	0.1	0.5

3. Attention | Wie kann das funktionieren? Wortvektoren! Schritt 2:

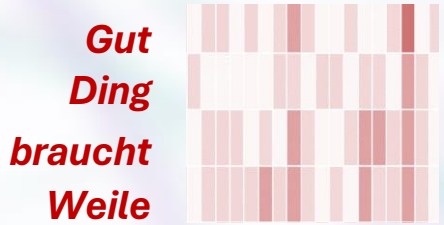


- Alle Input-Vektoren werden durch drei unterschiedliche Neuronale Netze gesendet.
- Es entstehen dabei die Q- die K- und die V-Vektoren.





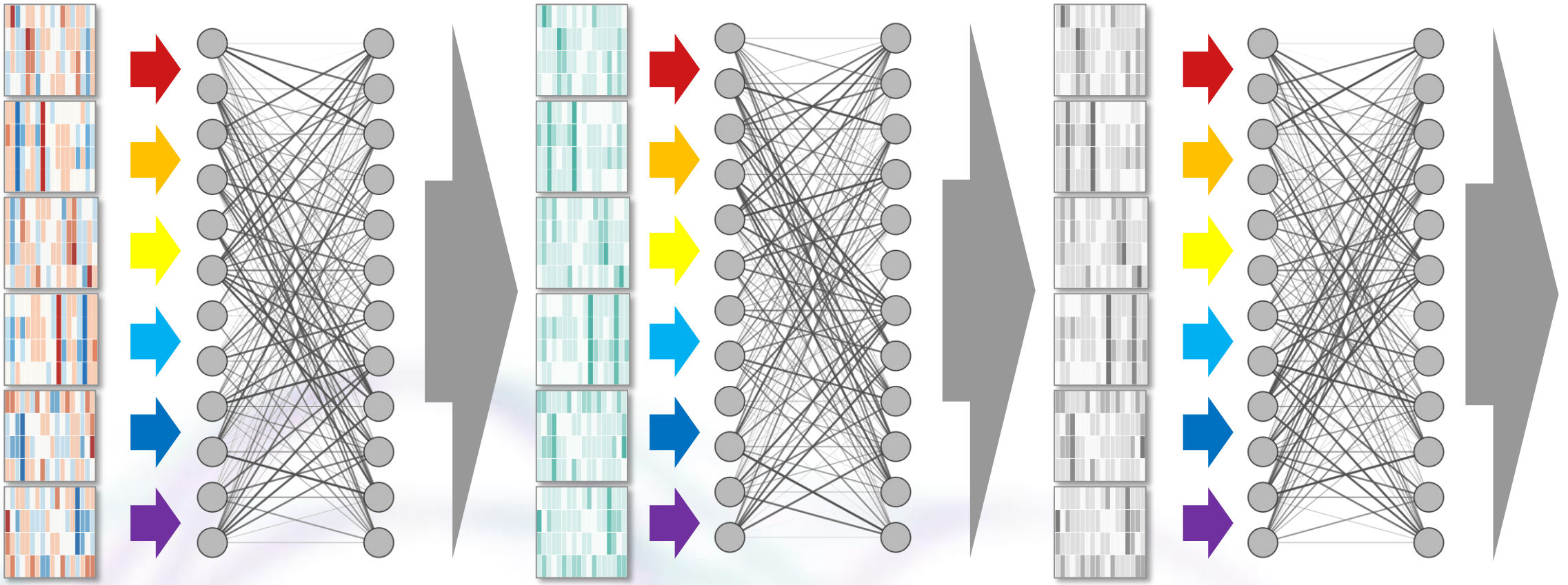
Query



Key

Value

3. Attention | Auf jeden Attention-Layer folgt ein Linear Layer



- Jeder Attention-Head HIER (!) verarbeitet 1/6 von d_{model}.
- Jeder Linear-Layer fügt alle 6 Einzel-Attentionscores wieder zu d_{model} zusammen.

4. GPT/Bard: Decoder-only | Start der Textgenerierung

- Die Attention Weights erzeugen eine Repräsentation des aktuellen Wortes.
- Diese Repräsentation wird in ein Feedforward-Netzwerk (FFN) eingespeist.
- Das FFN gibt eine Wahrscheinlichkeitsverteilung über mögliche nächste Wörter aus.

https://demo.allennlp.org/next-token-lm