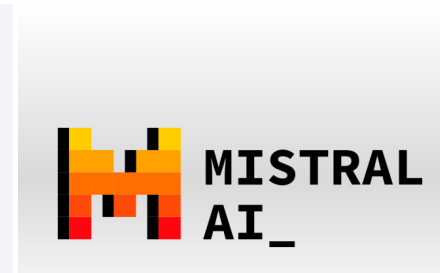




## Lokal installierte SLM / LLM ...

Thomas Jörg,  
LFTKI Baden-Württemberg,  
Kepler Gymnasium Pforzheim



deepinfra



LM Studio



Chatbox AI

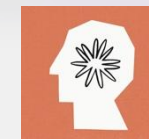


Jan



AnythingLLM

... versus ,Big Models‘:





## Lokal installierte SLM / LLM ...



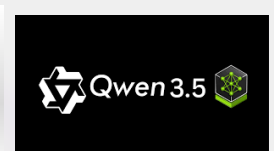
- Use Cases: Wo sind SLM / LLM nützlich?
- Welche rechtlichen Rahmenbedingungen für die Schule gibt es?
- Welche Hardware braucht man?
- Welche Software braucht man?
- Welche Eigenschaften haben diese Modelle?
- Wo liegen derzeit die Grenzen? Und wohin geht die Entwicklung?



## Die Handlungsempfehlung der Bildungsminister-Konferenz:

[https://www.kmk.org/fileadmin/veroeffentlichungen\\_beschluesse/2024/2024\\_10\\_10\\_Handlungsempfehlung-KI.pdf](https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2024/2024_10_10_Handlungsempfehlung-KI.pdf)

Das Dokument empfiehlt ausdrücklich, Schüler:innen im Umgang mit KI zu unterrichten. Der Unterricht über KI wird als eine neue Schlüsselkompetenz für das 21. Jahrhundert angesehen. Die Empfehlungen betonen, dass Schüler:innen durch den Unterricht in die Lage versetzt werden sollen, KI-Technologien kritisch und reflektiert anzuwenden. Dies soll ihnen helfen, als digitale Bürger:innen souverän zu agieren.



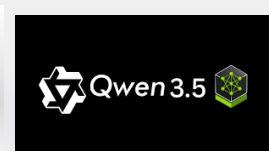


## Die Rechtlichen Rahmenbedingungen des EU AI Act:

<https://artificialintelligenceact.eu/>

seit August 2025 stufenweise in Kraft. Die wichtigsten Punkte:

- KI-Systeme werden in Risikostufen klassifiziert (minimal, begrenzt, hoch, unakzeptabel).
- Social Scoring und biometrische Massenüberwachung in Echtzeit sind verboten.
- KI-Systeme, die über Prüfungsergebnisse entscheiden, gelten als Hochrisiko-KI und unterliegen strengen Auflagen.
- Transparenzpflicht: Nutzer müssen informiert werden, wenn sie mit einer KI interagieren.





## Warum nicht GPT5.4 / Gemini 3.1 / Claude 4.6, die sehr leistungsfähig sind?

- Bei den ‚Big Three‘ (GPT/Claude/Gemini) problematisch: Datenschutz / Privacy
- Für cloudbasierte Modelle ist immer eine Internetverbindung nötig.
- Für viele Aufgaben ist das umfassende Wissen der ‚Großen‘ nicht unbedingt nötig
- Leistungssprung der SLMs: Modelle wie Qwen3-8B oder Phi-4 erreichen 2026 ein
- Qualitätsniveau, das vor 18 Monaten nur die „Big Three“ liefern konnten.
- EU AI Act verstärkt Datenschutzpflichten: Der regulatorische Druck auf US-Cloud-Dienste im europäischen Bildungssektor wächst weiter.





Warum nicht GPT / Gemini / Claude ...

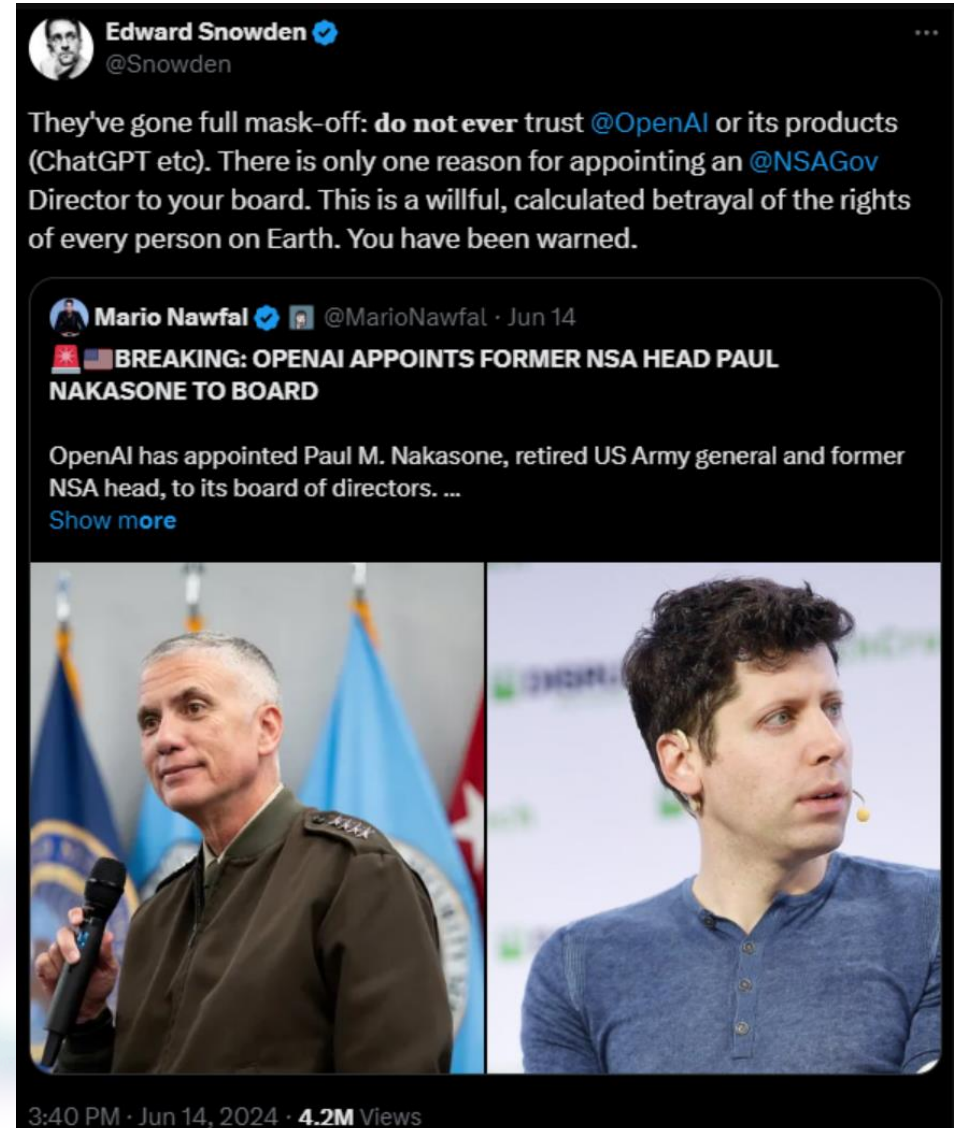


*„He concurrently served as the director of the National Security Agency NSA“*



*„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-verbundung-fuer-nicht-mehr-vertrauenswuerdig/>





**Welche Lösungsansätze gibt es bereits?** *Und welche Gegenargumente?*

- Fobizz gilt als datenschutzrechtlich problematisch (z.B. Aussage LMZ Reutlingen):  
*„Wir hatten eine Fobizz-Lizenz gekauft, die wir aber nicht weiter verleihen dürfen.“*
- Dazu die FoBi „Datenschutz und KI in Bildungseinrichtungen“:  
<https://www.baden-wuerttemberg.datenschutz.de/offene-veranstaltung-2024-ps-803-buchen/>  
<https://www.baden-wuerttemberg.datenschutz.de/wp-content/uploads/2024/07/2024-PS-803-KI-Ausschreibung-HP.pdf>
- ABER, bisher gängige Meinung:  
[https://fg-freiburg.de/fg-wAssets/docs/support/2024\\_Fobizz\\_Datenschutz\\_KI.pdf](https://fg-freiburg.de/fg-wAssets/docs/support/2024_Fobizz_Datenschutz_KI.pdf)





## Teil 1: Was ist erlaubt und gewünscht? Stand 2024

*„Die Schule hat einen eigenen Zugang, API-Schlüssel und AVV mit dem Anbieter abgeschlossen und erstellt mit Oberstufenschülern eine eigene Plattform zur Nutzung der KI via API. Sind die Spielregeln klar, ist eine Nutzung mit älteren Schülerinnen und Schülern ab 16 Jahren möglich. Die Risiken sind begrenzt.*

*Die Schule hat lokal als App auf Computern oder Tablets laufende KI-Anwendungen. Eine Nutzung ist für Schülerinnen und Schüler ohne Risiken möglich, da keine Daten an Dritte abfließen. Trotzdem sollte es zuvor abgesprochene Spielregeln geben.“*

<https://datenschutz-schule.info/tag/chatgpt/>



## KI in Prüfungen - KMK-Handlungsempfehlung Stand 2026

- *Kein generelles KI-Verbot bei Hausarbeiten*
- *Neue Prüfungsformate: Präsentation + Verteidigung, Prozessbewertung*
- *4K-Kompetenzen im Fokus: Kommunikation, Kollaboration, Kreativität, Kritisches Denken*
- *Kennzeichnungspflicht für KI-generierte Inhalte*
- *Mehrdimensionale Bewertung (schriftlich, mündlich, praktisch, digital)*

[https://www.kmk.org/fileadmin/veroeffentlichungen\\_beschluesse/2024/2024\\_10\\_10-Handlungsempfehlung-KI.pdf](https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2024/2024_10_10-Handlungsempfehlung-KI.pdf)



## Was kann man mit lokalen Sprachmodellen (sonst noch) machen?

- **RAG: „Retrieval-Augmented Generation“**

Problem von Halluzinationen & beschränktem Erfahrungsschatz wird gelöst: LLMs werten Informationsquellen aus.

- **Agentic Systems**

ReAct (Reasoning-Acting): Sprachmodelle als autonome Planungsinstrumente, orchestrieren weitere LLMs und Tools/Skills, die Aufgaben erledigen.

- **Innovatoren in der Forschung**

erste Anzeichen bei AlphaFold 3, mathematischen Beweisen.

*We are here*



**Level 1:** Chatbots

**Level 2:** Reasoner, können Logik anwenden und Probleme lösen

**Level 3:** Agenten, können zusätzliche Aktionen ausführen

**Level 4:** Innovatoren, können neue Erfindungen machen

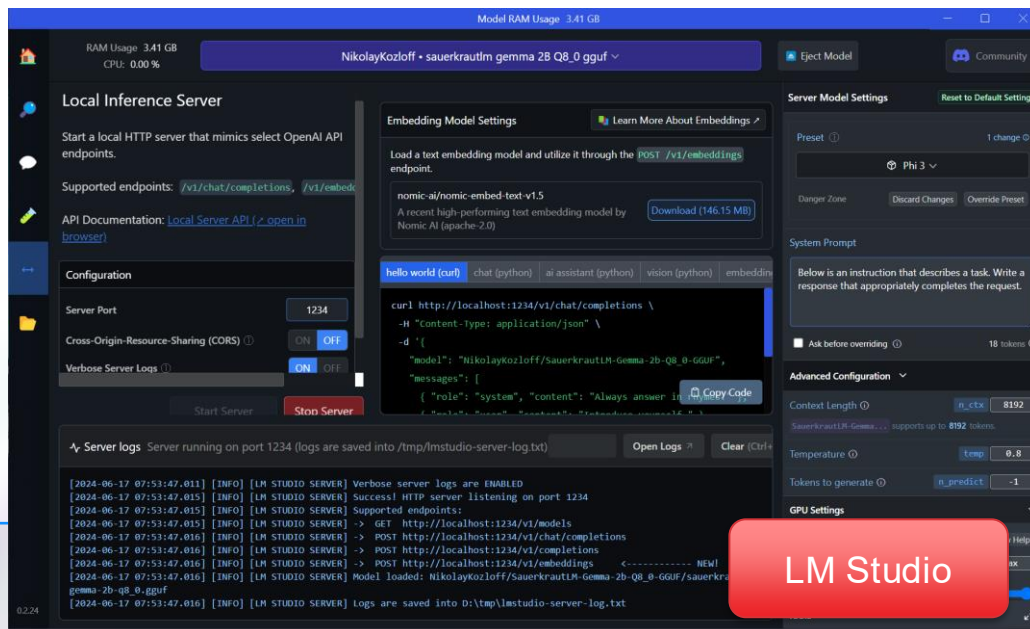
**Level 5:** Kann die Arbeit einer ganzen Organisation erledigen

<https://openai.com/charte/>

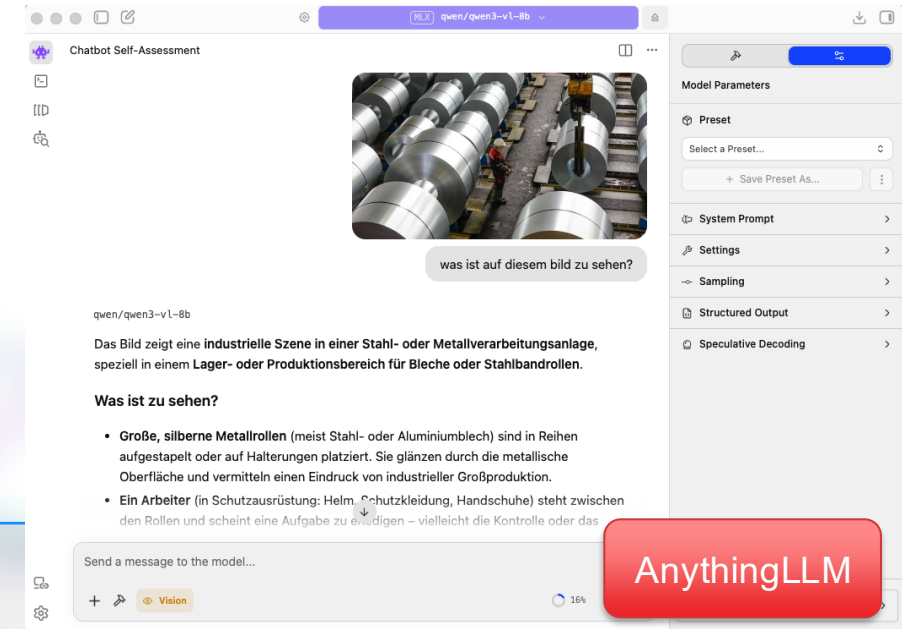


## Was funktioniert ,jetzt‘? (9. März 2026) *Ein Überblick*

- **Opensource-LLMs** (*werden auch von Apple/Microsoft verwendet*)
- **Software, um LLMs/SLMs zu hosten** (*z.B. LM Studio, Ollama*)
- **Software für RAG & Agents** (*z.B. AnythingLLM, OpenClaw*)



<https://www.lmstudio.ai>

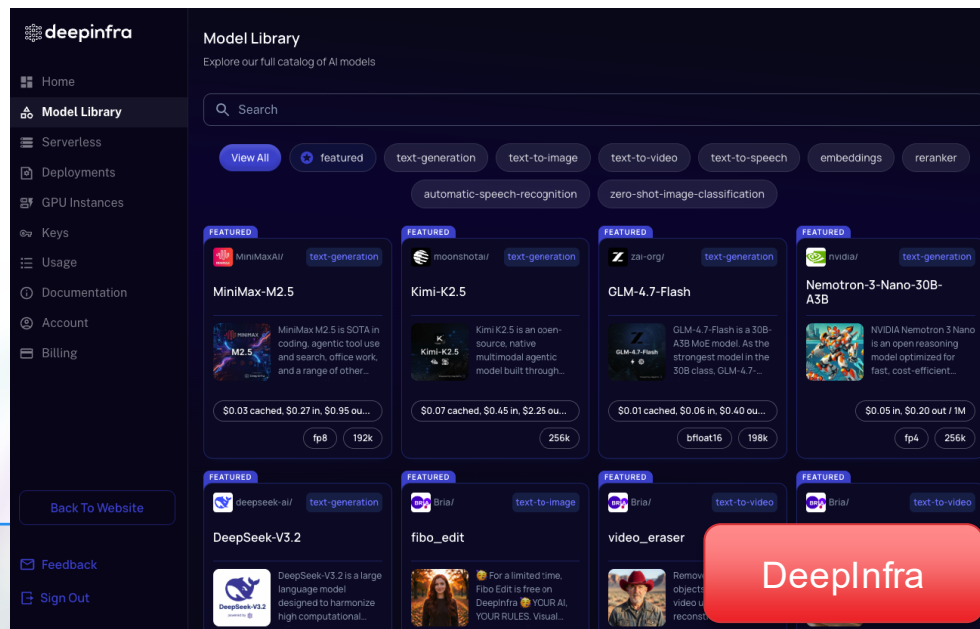


<https://anythingllm.com>

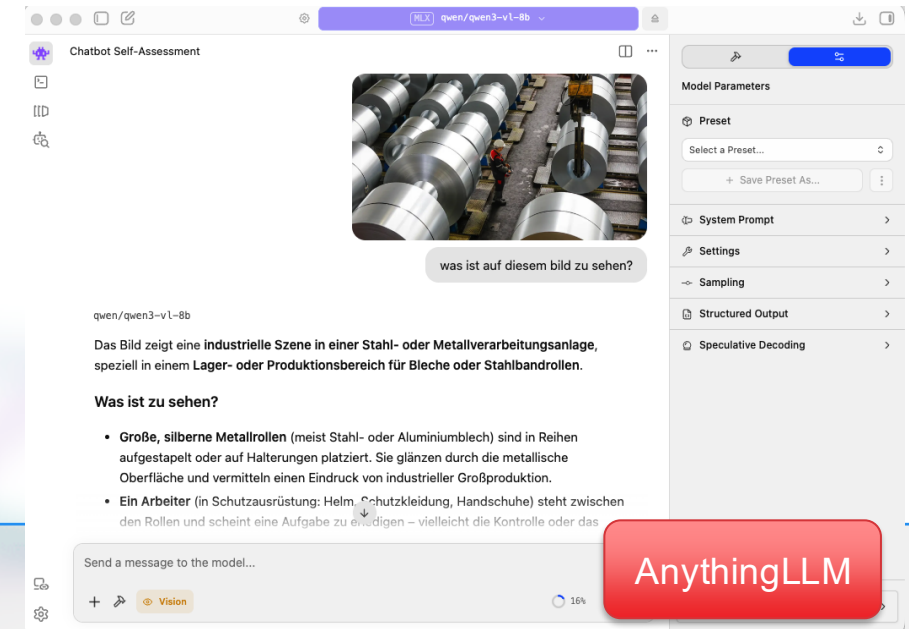


## Was funktioniert ‚jetzt‘? (9. März 2026) Ein Überblick

- Große Modelle selbst hosten auf DeepInfra: *Deepseek V3.2 ist vollständig anonym.*
- + Frontend für LLM auf Schulrechnern: *AnythingLLM als (RAG-) Client.*
- Überschaubare Kosten (vgl. Fobizz): *1 \$ für 1 Million Token (ca. 750.000 Wörter)*



<https://deepinfra.com/>



<https://anythingllm.com/>



## Was funktioniert ,jetzt'? (15. Oktober 2024) *Im Einzelnen: Lokale Installationen*

### DEMO LM Studio & Anything LLM

- LM Studio (SLM Hosting):

<https://lmstudio.ai/>

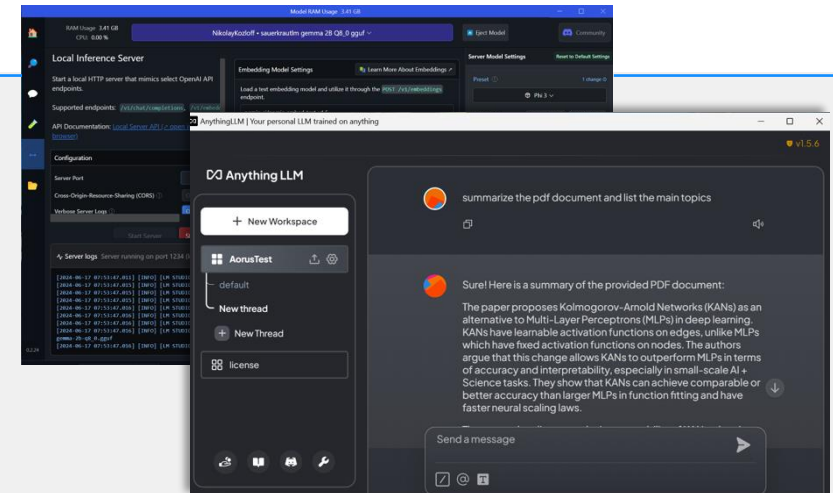
- Anything LM (RAG):

<https://useanything.com/>

- Huggingface: 😊

*"riesiger Basar fuer KI"*

[https://huggingface.co/models?pipeline\\_tag=text-generation&sort=downloads](https://huggingface.co/models?pipeline_tag=text-generation&sort=downloads)





Was funktioniert ,jetzt'? (15. Oktober 2024) *Im Einzelnen: Lokale Installationen*

## DEMO LM Studio & Anything LLM

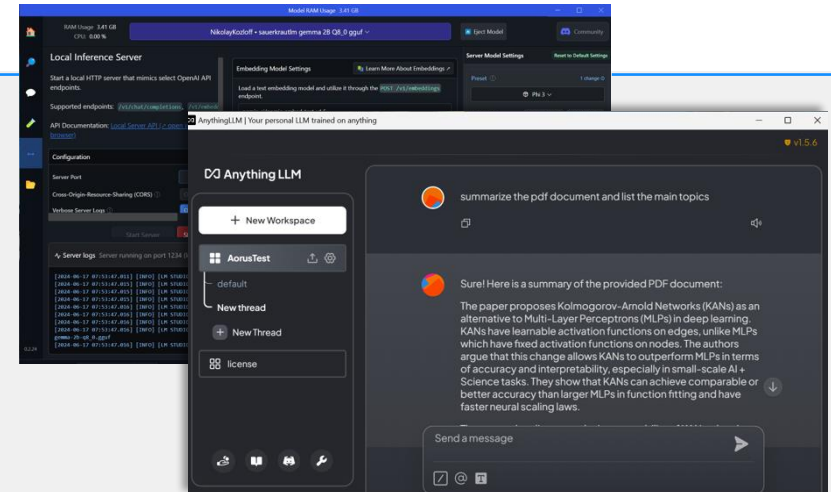
- LM Studio:

*Suche, Herunterladen, Grafikkarten-Offload,  
Server. Achtung: **Context length!***

- Anything LM:

*General Setup (**Token Context Window!**), Workspace, Document Database,  
Embeddings, Whisper, **Agent Configuration!***

*sollten gleich sein.*





## Was funktioniert ‚jetzt‘? (15. Oktober 2024) *Im Einzelnen: Große Modelle selbst hosten*

### DEMO Jan.ai & DeepInfra

- <https://deepinfra.com/>



deepinfra

**Stellt Serverhardware UND  
OpenSource LLMs zur Verfügung**

| Model                                   | Type            | Usage             | Rate           | Spend         |
|-----------------------------------------|-----------------|-------------------|----------------|---------------|
| meta-llama/Meta-Llama-3.1-405B-Instruct | text-generation | 34,025 out tokens | \$17900/Mtoken | \$0.06        |
| <b>Total*</b>                           |                 |                   |                | <b>\$0.31</b> |

| September 2024                                          |                 |                   |                 |               |
|---------------------------------------------------------|-----------------|-------------------|-----------------|---------------|
| Model                                                   | Type            | Usage             | Rate            | Spend         |
| meta-llama/Meta-Llama-3.1-405B-Instruct from 2024-09-03 | text-generation | 167012 in tokens  | \$17900/Mtoken  | \$0.30        |
| meta-llama/Meta-Llama-3.1-405B-Instruct from 2024-09-03 | text-generation | 50,512 out tokens | \$17900/Mtoken  | \$0.09        |
| meta-llama/Meta-Llama-3.1-70B-Instruct                  | text-generation | 7,254 in tokens   | \$0.3500/Mtoken | \$0.00        |
| meta-llama/Meta-Llama-3.1-70B-Instruct                  | text-generation | 3,865 out tokens  | \$0.4000/Mtoken | \$0.00        |
| <b>Total*</b>                                           |                 |                   |                 | <b>\$0.39</b> |

- <https://anythingllm.com>



Jan

**Frontend für LLMs, kann auch lokale SLM einbinden. Inklusive RAG**



## Gibt's sonst noch was neben DeepInfra oder Jan.ai?

- <https://openrouter.ai> Hub für viele Anbieter, bietet API-Keys & Skriptzugriff, Mögliches Problem: DSGVO, weil viele Anbieter unter einem Dach. Problematisch: Openrouter gibt zu, dass sie teilweise (vom Anbieter abhängig) den Datenverkehr mitschneiden. Zwar ist die Anonymisierung im Schulkontext nach wie vor gegeben, aber: Es geht mit Deepinfra / MistralAI besser!

- <https://mistral.ai>  
[„La Plateforme“ bietet API-Keys und Skriptzugriff]

**europäisch! 😊**

<https://console.mistral.ai/home>





## Welche online gehosteten Modelle betrachten wir in diesem Vortrag?

- Llama 3.1 405B (bei Deepinfra). Völlig unklar sind die rechtlichen Rahmenbedingungen

*„Das Oberlandesgericht Karlsruhe hat ein wegweisendes Urteil zum Datenschutz (auch) an Bildungseinrichtungen gefällt. Öffentliche Auftraggeber, also auch Schulen und Schulträger, können darauf vertrauen, wenn ihnen IT-Anbieter Datenschutz-Kompatibilität zusichern – und sie beauftragen.“*

<https://www.news4teachers.de/2022/09/wegweisendes-urteil-behoerden-schulen-duerfen-darauf-vertrauen-wenn-it-anbieter-ihnen-datenschutz-kompatibilitaet-zusichern/>

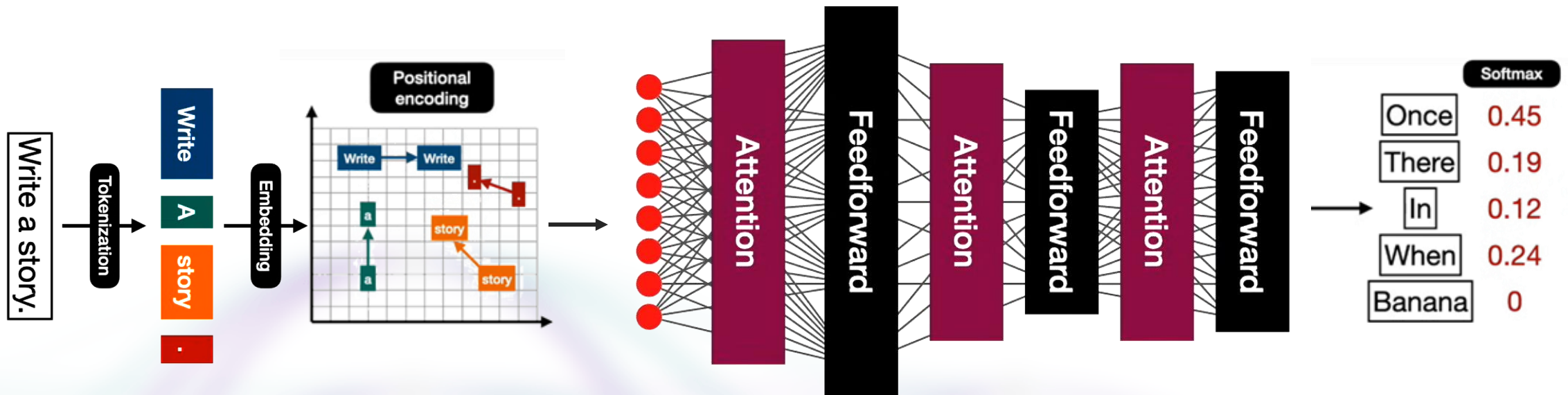
*„Auch wenn die großen US Anbieter in ihren Datenschutzerklärungen [...] mittlerweile ausschließen, von Nutzern eingegebene Prompts und andere Daten zu Trainingszwecken zu verwenden, bleibt die Übermittlung von personenbezogenen Daten auf die Server von US Anbietern problematisch und ist im Rahmen unterrichtlichen Nutzung von KI Plattformen zur Erfüllung des Bildungs- und Erziehungsauftrags von Schulen nicht vertretbar.“*

<https://datenschutz-schule.info/tag/chatgpt/>



## Grundlagen: Was ist ein LLM (Large Language Model)? Grundaufbau LLM

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



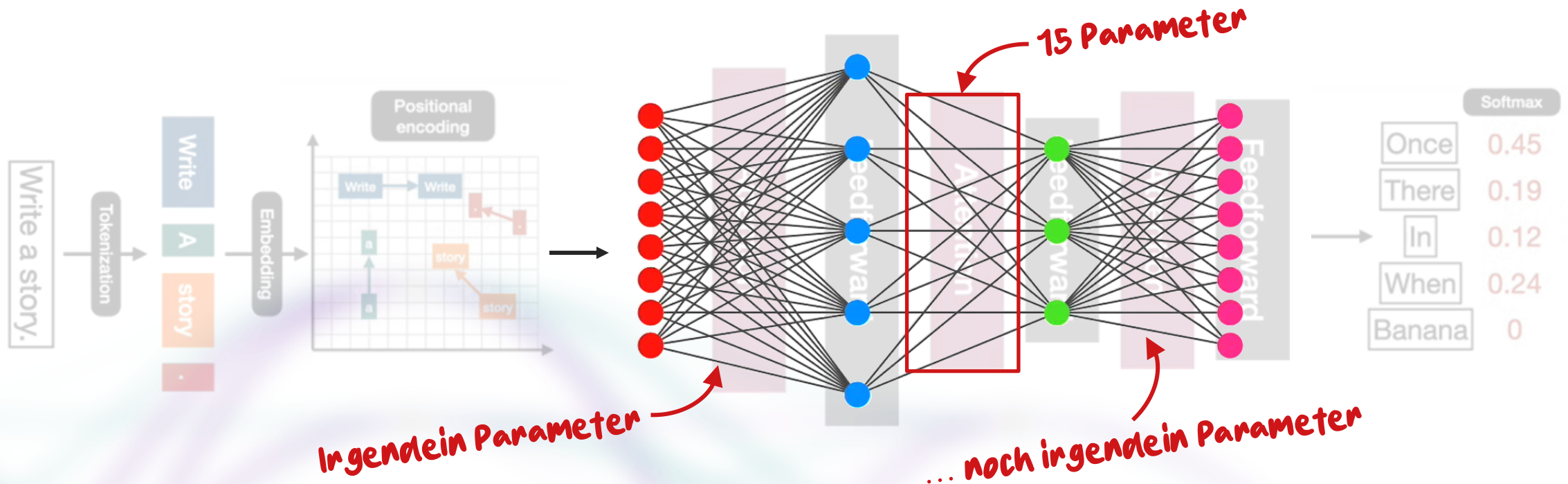
What are Transformer Models and how do they work? (Luis Serrano)

<https://www.youtube.com/watch?v=qaWMOYf4ri8>



## Grundlagen: Was ist ein LLM (Large Language Model)? Parameter eines LLM

- 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>



## Grundlagen: Was ist ein Attention-Mechanismus? Funktionsprinzip Attention-Head

„Gut Ding braucht Weile“

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

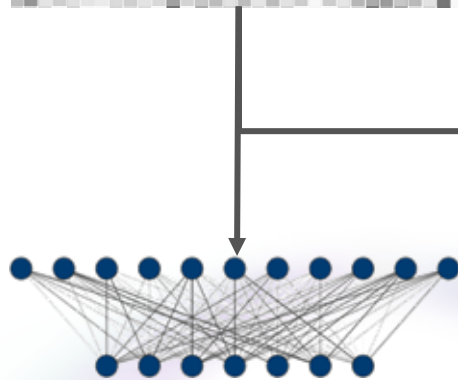
|              | KEY<br>Gut | Ding       | braucht    | Weile      |
|--------------|------------|------------|------------|------------|
| QUERY<br>Gut | 0.3        | <b>0.5</b> | 0.1        | 0.1        |
| Ding         | 0.2        | 0.3        | <b>0.4</b> | 0.1        |
| braucht      | 0.2        | 0.1        | 0.2        | <b>0.5</b> |
| Weile        | 0.1        | 0.3        | 0.1        | 0.5        |



## Grundlagen: Was ist ein Attention-Mechanismus? Vektorielle Darstellung

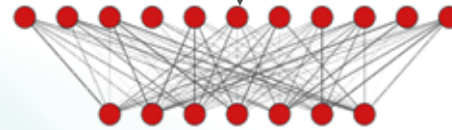
Gut  
Ding  
braucht  
Weile

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



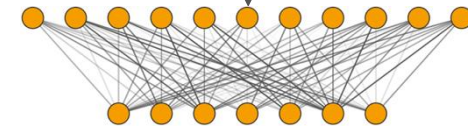
Gut  
Ding  
braucht  
Weile

**Query**



Gut  
Ding  
braucht  
Weile

**Key**

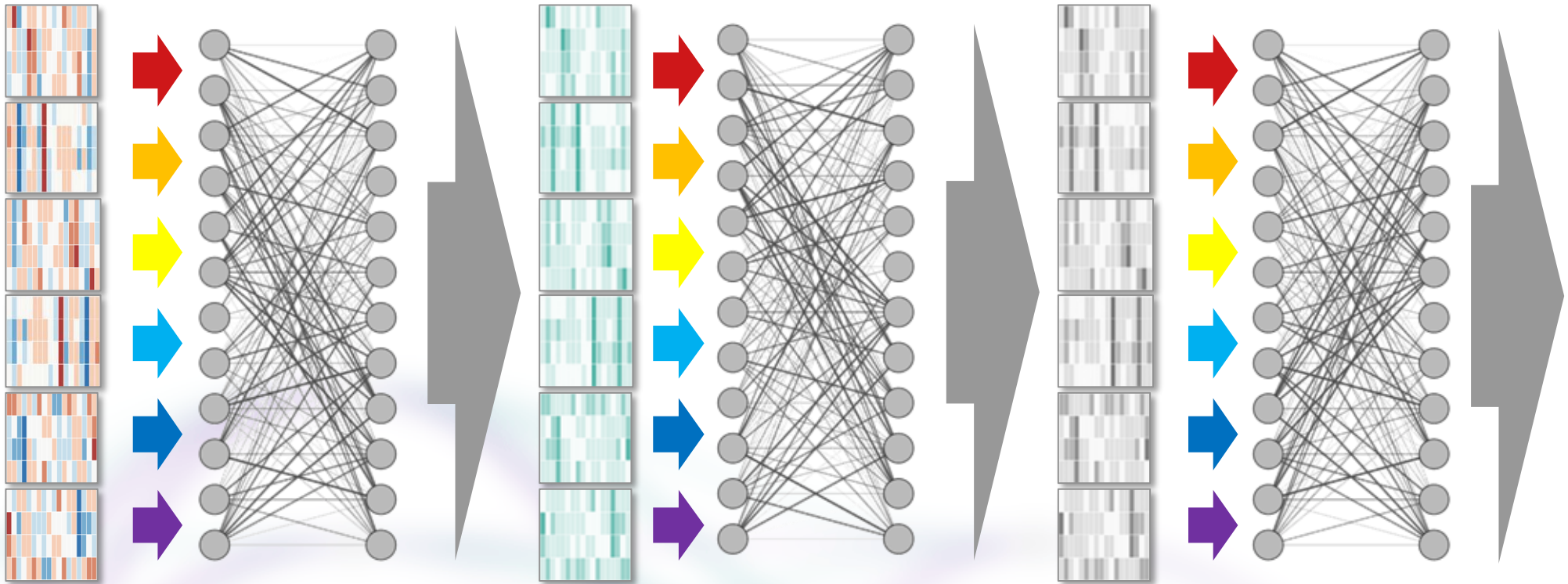


Gut  
Ding  
braucht  
Weile

**Value**



## Grundlagen: Was sind Attention-Layer? Schematischer Aufbau



- In diesem Beispiel: Jeder Attention-Head verarbeitet 1/6 der Information eines Wortes.
- Jeder Linear-Layer fügt alle 6 Einzelinformationen jedes Attention-Heads zusammen.



## Welche Hardware-Vorraussetzungen gibt es? *Teil 1 von 4*

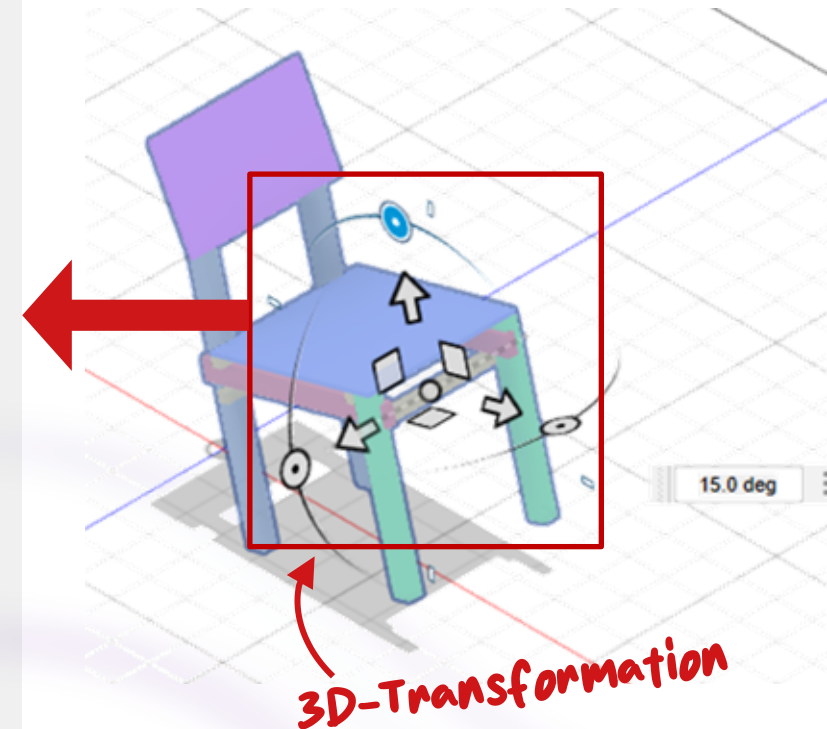
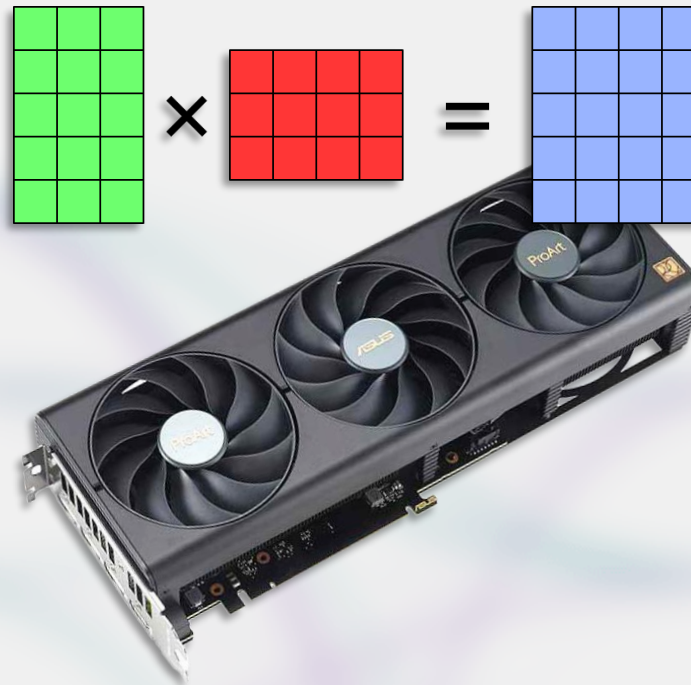
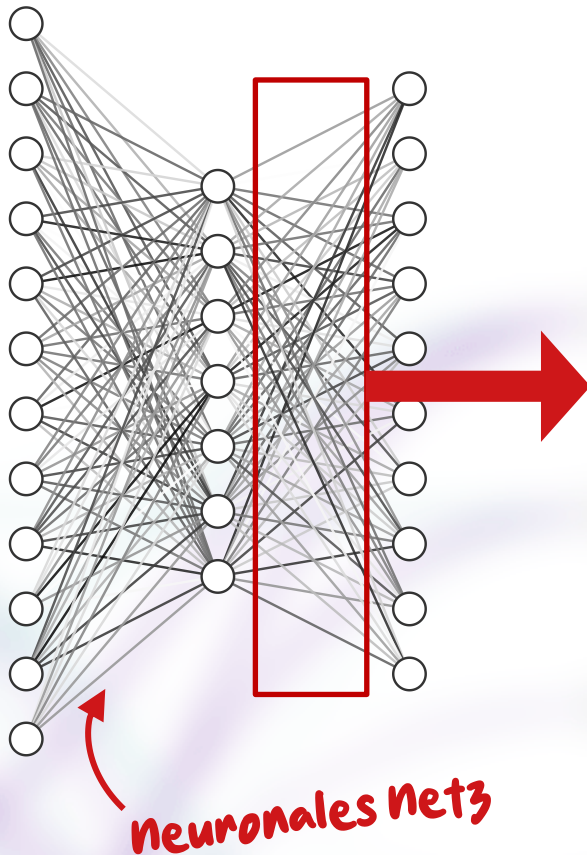
- „Das natürliche Biotop von neuronalen Netzen sind GPU und NPU/TPU“.
- Glücklicherweise sind neuronale Netze auf NVIDIA-Gaming-GPUs lauffähig.
- UND: in immer stärkeren Maß auf Apple Silicon (M1 ... M5)!  
Apple nutzt ‚Unified Memory‘; LLMs werden ins normale RAM ausgelagert. Vorteilhaft!
- *Stand Frühjahr 2026: Noch NVIDIA, immer mehr auf Apple M4 / M5*
- *Gescheiterter Hoffnungsträger: Snapdragon mit integrierter NPU*





## Welche Hardware-Vorraussetzungen gibt es? *Teil 2 von 4*

- Warum GPU? *Beide Bereiche nutzen parallelisierte Matrizenrechnungen*





## Welche Hardware-Vorraussetzungen gibt es? *Teil 3 von 4*

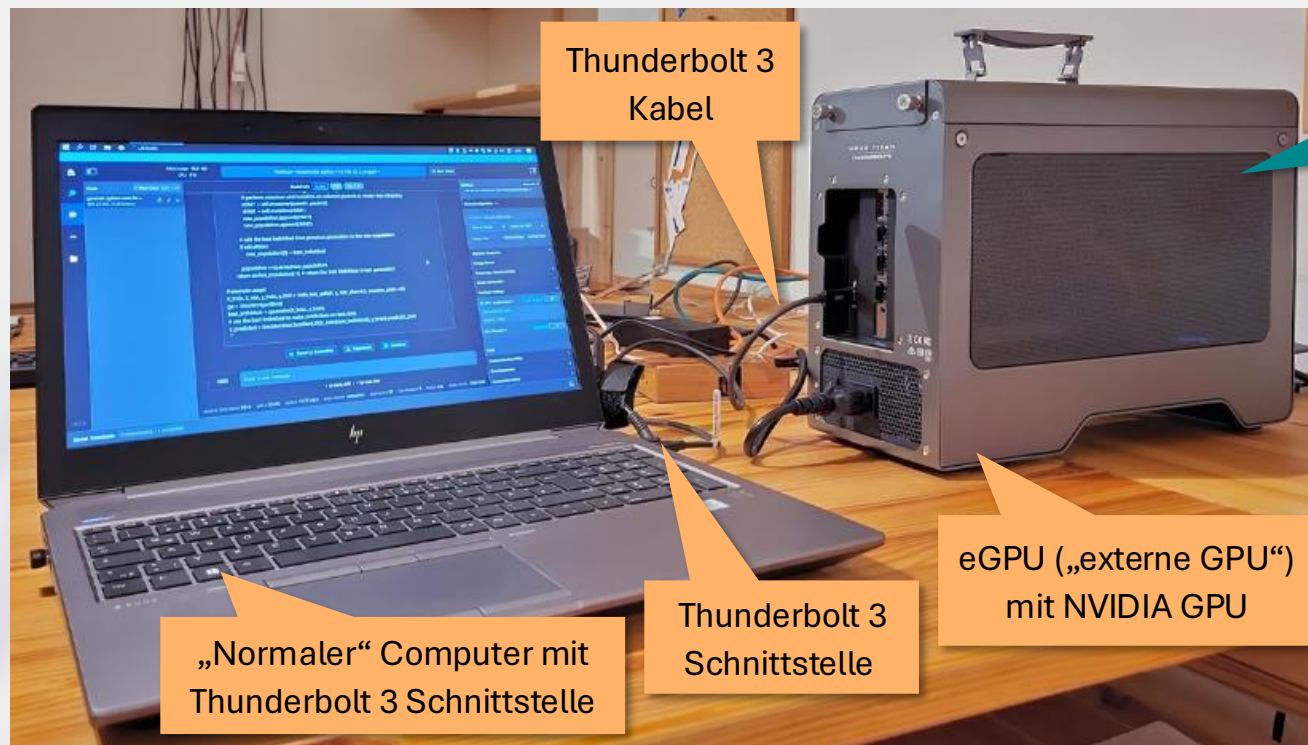
- VRAM (bei NVIDIA) und ‚Unified RAM‘ bei Apple / Microsoft
- Je mehr desto besser. Aber mindestens (!! ) **12GB** dediziert für NPU / GPU.
- Obergrenze derzeit bei NVIDIA: 24 GB Desktop (4090), 16 GB Laptop (3080 / 3080Ti)
- Obergrenze derzeit bei Apple: Geldbeutel, bzw. 256GB
- *Obergrenze derzeit bei Snapdragon: bei ,0‘ es existiert immer noch kein Framework!*





## Welche Hardware-Vorraussetzungen gibt es? *Teil 4 von 4*

- Egpu mit externer Grafikkarte über Thunderbolt-3 Schnittstelle an Rechner.
- Empfehlung: 16GB Nvidia-GPUs oder 24GB Nvidia-GPUs, je nach LLM & Use Case



Akitio Node Titan  
<https://www.akitio.com/expansion/node-titan>

Informationsquellen zu EGPUs:

<https://egpu.de/>

<https://egpu.io/>

<https://geizhals.de/?cat=hwegpu>



## Welche Hardware-Vorraussetzungen gibt es? *Teil 4 von 4*

- Egpu mit externer Grafikkarte über Thunderbolt-3 Schnittstelle an Rechner.
- Von uns getestet (Unterricht & zuhause): Sonnet Breakaway Box

RTX 3090 24GB OC  
ca. 1500 Euro

RTX 4060Ti 16GB OC  
Ca. 450 Euro



Sonnet Breakaway Box 750  
<https://sonnettech.com/product/egpu-breakaway-box/>

ca. 450 Euro

### Kompatible GPUs:

[https://www.sonnettech.com/support/downloads/manuals/Compatibility\\_Graphics\\_Cards.pdf](https://www.sonnettech.com/support/downloads/manuals/Compatibility_Graphics_Cards.pdf)



Last but not least: Warum hosten wir nicht selbst z.B. ein Llama 405B? 🤔

- 191 Files mit jeweils 4,5GB sind circa **900 GB** Dateigröße.  
<https://huggingface.co/meta-llama/Llama-3.1-405B/tree/main>
- Diese Datei muss sich vollständig im Grafikkarten-Speicher befinden (Vram).
- Dazu sind mindestens 12 NVIDIA H100 nötig (80GB Vram pro Stück)
- Diese kosten pro Stück (Stand 15. Oktober 2024): 37.000 Euro  
<https://geizhals.de/nvidia-h100-900-21010-0000-000-a3264418.html>
- Ein dedizierter Server benötigt eine Stromversorgung von mindestens 3,8 kW (!)  
<https://www.dell.com/de-de/dt/servers/specialty-servers/poweredge-xe-servers.htm#tab0=0&accordion0>
- ... und ganz erhebliches besonderes Know How  
(ich möchte nicht derjenige sein, der einen Server mit ca. 500.000 Euro Anschaffungskosten als erster ‚anschaltet‘ ...)



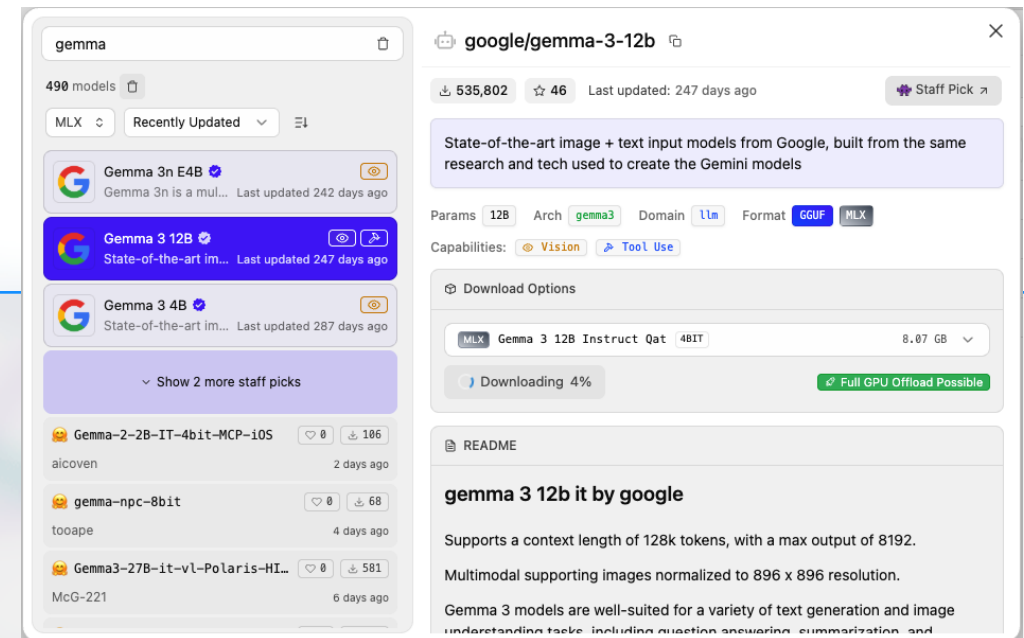


## Konkret: Welche OpenSource Modelle sind (9. März 2026) besonders nützlich?

- **Small-Medium 7B, 8B:** Qwen 3-8B, Qwen 3.5-9B
- **Medium: 12B, 20B:** Gemma 3-12B, GPT-OSS-20B
- **FIM:** Qwen 2.5 Coder-14B,  
Mistral Codestral, DeepSeek 3.2

**Wichtig mittlerweile:**

**Die Modelle sollten trainiert sein für  
,Tool-Use‘**





## Das nächste Problem: Modellgrößen

|     |                            |                                              |          |
|-----|----------------------------|----------------------------------------------|----------|
| F32 | Meta Llama 3.1 8B Instruct |                                              | 32.13 GB |
| F32 | Mistral Nemo Instruct 2407 | × Wahrscheinlich zu groß für diesen Computer | 49.00 GB |

*Floating Point 32 Bit!*



- FP 16 Modelle sind exakt halb so groß. Also immer noch zu groß für normales Vram.
- Man benötigt eine Art ‚Komprimierung‘, ähnlich wie bei .jpg / .mp3 / .gif
- Also ein verlustbehaftetes Verwerfen von Bitauflösung:

**Quantisierung ist die Lösung des Problems!**

*(für eventuelles Nachtrainieren / Feintuning von LLMs benötigt man FP32)*



Welche Quantisierungen sind für ein lokales Modell vernünftig? Merke:

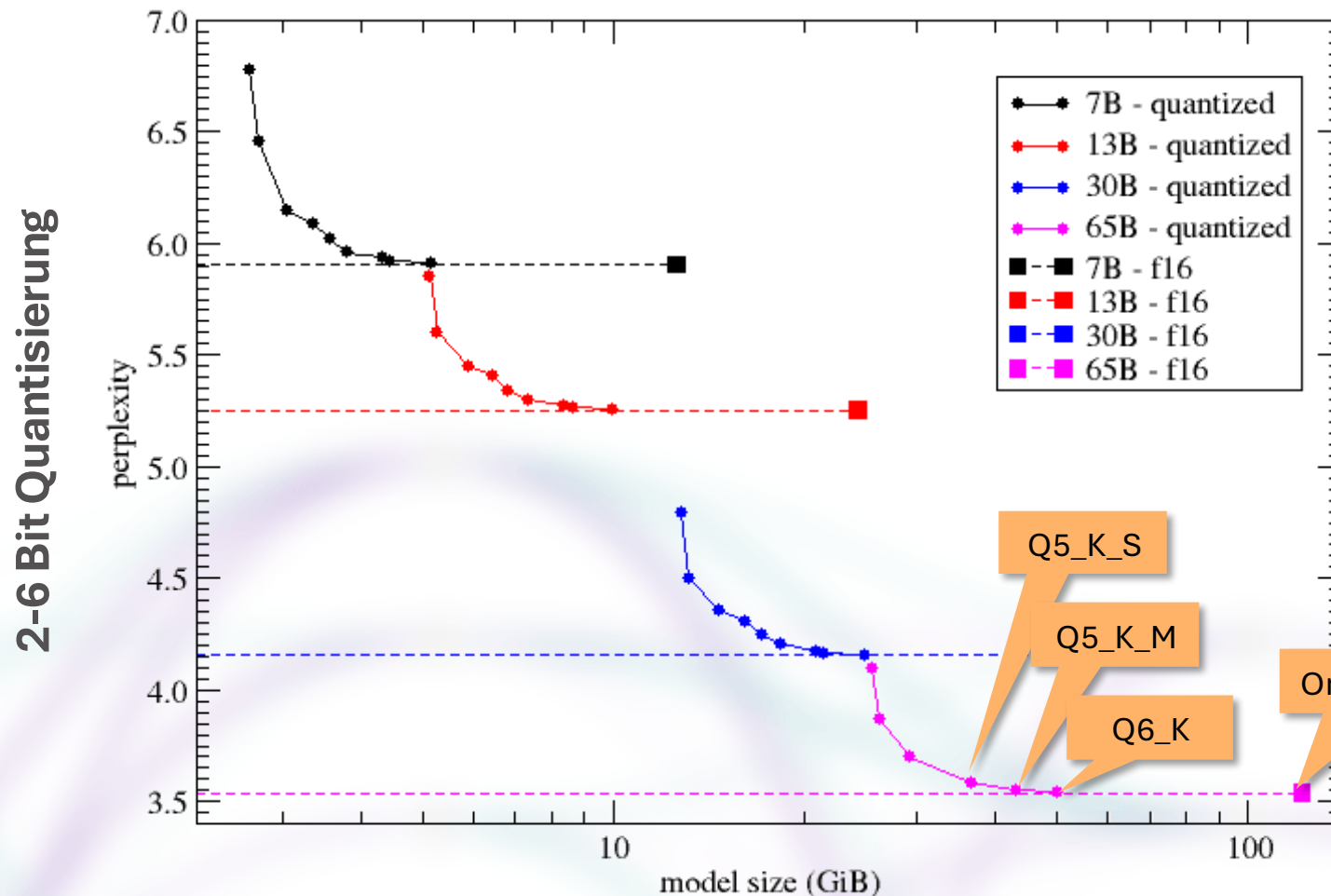
- Q8\_0 | Q6\_K | Q5\_K\_M 🧐👉
- <https://huggingface.co/dranger003/c4ai-command-r-plus-iMat.GGUF>  
<https://github.com/gggerganov/llama.cpp#quantization>  
<https://github.com/gggerganov/llama.cpp/pull/1684#issue-1739619305>

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?



“A lower-weight quant of a larger model will always have better perplexity than even an unquantized smaller model, except that 2-bit quants approach perplexity similar to a model of half the parameter count.”

<https://rentry.org/llama-cpp-quants-or-fine-ill-do-it-myself-then-pt-2>

<https://github.com/ggerganov/llama.cpp/pull/1684#issue-1739619305>



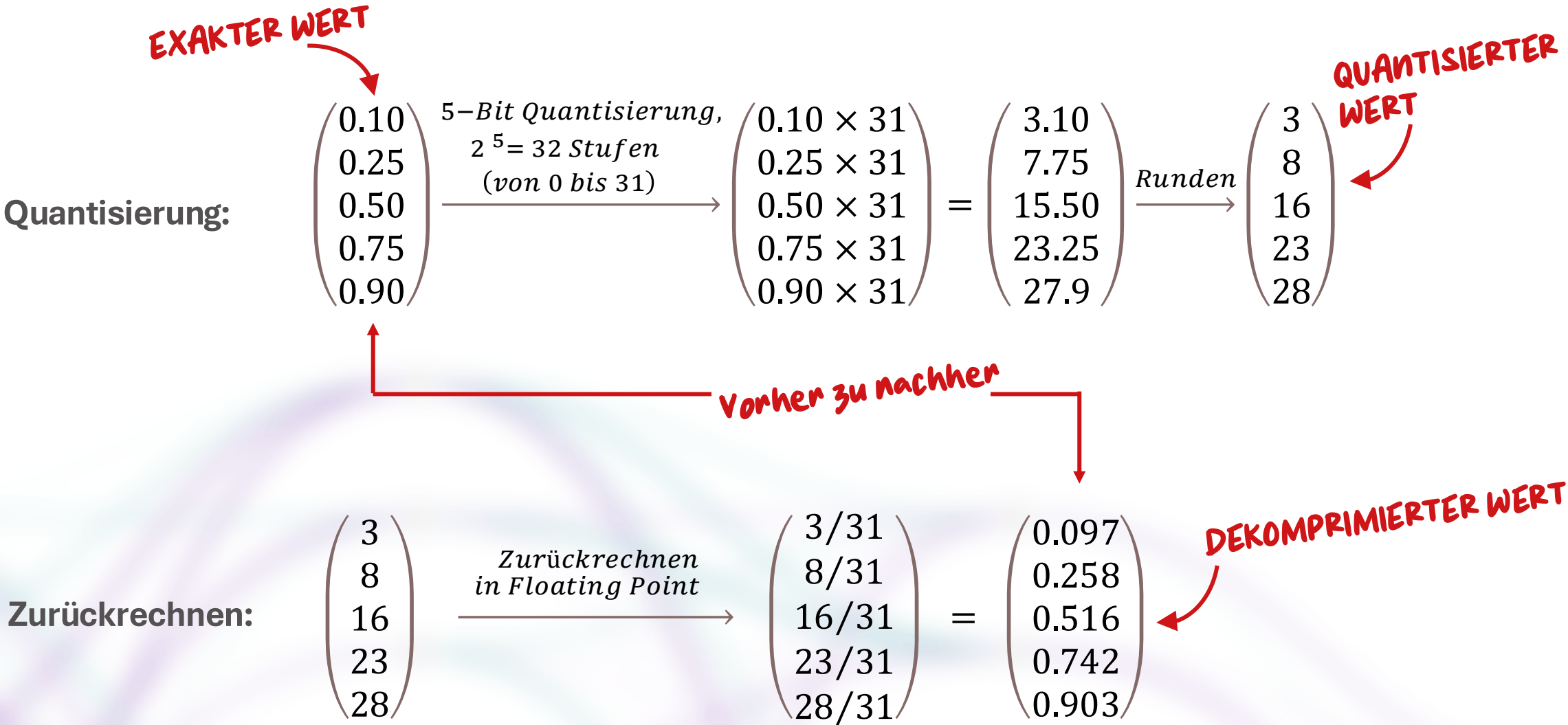
### Welche Quantisierungen sind für ein lokales Modell vernünftig?

| 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,<br>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.



# Wie berechnet man die Quantisierung? Beispiel Q5\_K\_M:





## Hardware-Setup: Wie kalkuliert man den Vram-Speicherbedarf eines SLM?

Faustregeln für ein 13B-Modell (Q5KM) & 10 Clients (grobe Abschätzung!):

- Modellgröße des quantisierten Modells: z.B. **9 GB**
- Activation Overhead ca. 5-10% des GPU-Vrams: **2 GB**
- + KV-Cache („Key-Value-Vectorcache“): **500 kB pro Token**
- Für 2048 Tokens Context-Window sind das: **1 GB pro Sequenz**
- Für 10 User sind das: **10 GB**
- Insgesamt, also Modell plus Cache:  **$9 + 2 + 10 \text{ GB} = \underline{21 \text{ GB}}$**

<https://ai.gopubby.com/stop-guessing-heres-how-much-gpu-memory-you-really-need-for-llms-8e9b02bcdb62>

<https://training.continuumlabs.ai/inference/why-is-inference-important/key-value-cache>

<https://kipp.ly/transformer-inference-arithmetic/>



**RTX 3090**



## Hardware-Setup: Wie kalkuliert man den VRAM-Speicherbedarf eines SLM?

Faustregeln für ein 8B-Modell (Q5KM) & 1 Client (grobe Abschätzung!):

- Modellgröße des quantisierten Modells: z.B. **6 GB**
- Activation Overhead ca. 5-10% des GPU-Vrams: **2 GB**
- + KV-Cache (,Key-Value-Vectorcache‘): **500 kB pro Token**
- Für 4096 Tokens Context-Window sind das: **2 GB pro Sequenz**
- Insgesamt, also Modell plus Cache: **6 + 2 + 2 GB = 9 GB**
- *Kommt eventuell noch RAG dazu, werden es etwa 4 GB mehr*

<https://ai.gopubby.com/stop-guessing-heres-how-much-gpu-memory-you-really-need-for-llms-8e9b02bcd62>

<https://training.continuumlabs.ai/inference/why-is-inference-important/key-value-cache>

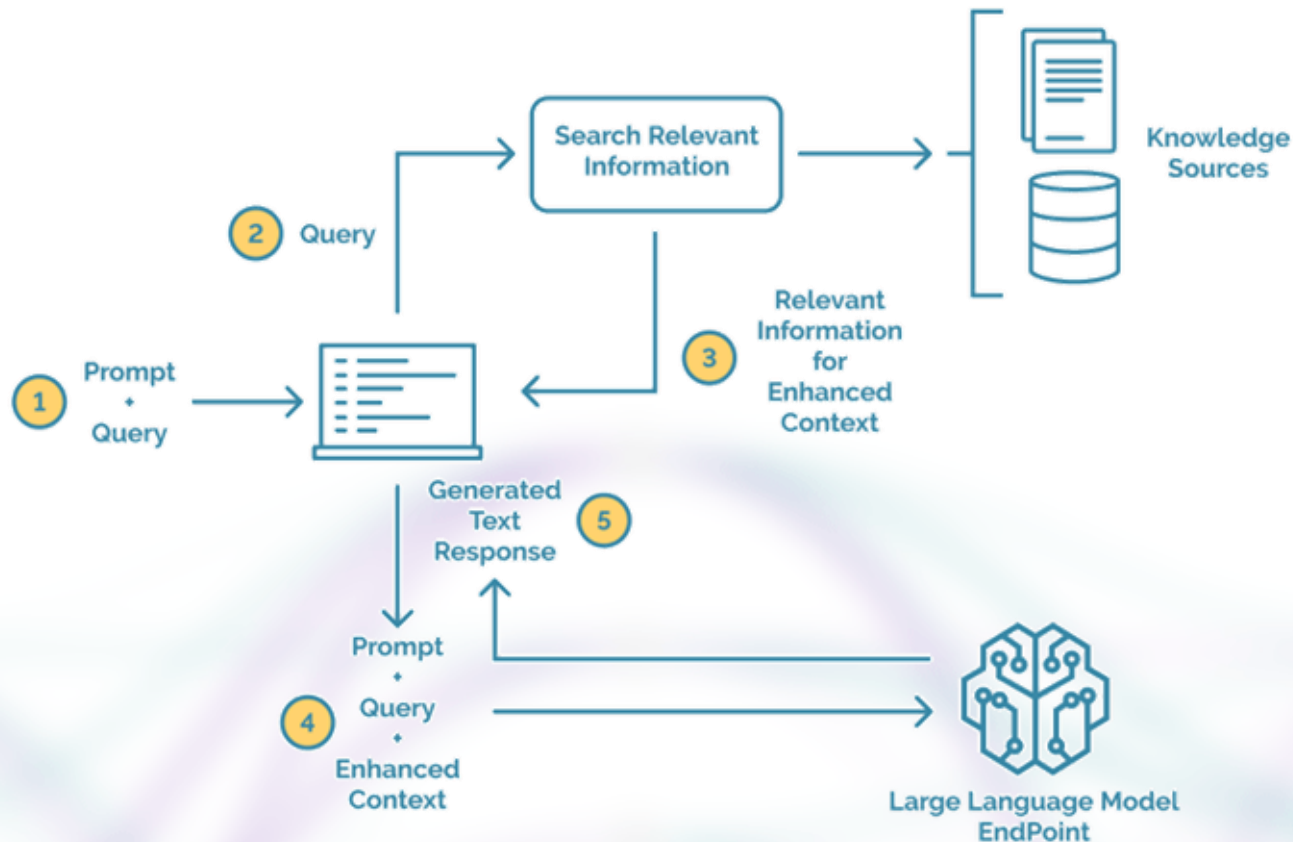
<https://kipp.ly/transformer-inference-arithmetic/>



**RTX 4060**



## RAG mit AnythingLLM: Setup und generelle Funktion



1. User stellt mittels Prompts eine Frage.
2. Ein Embedding-Modell sucht in einer Vektordatenbank in vorher erstellten Document-Embeddings nach relevanten Document Chunks,
3. ... und liefert solche Chunks zurück, deren Embeddings mit der Query eine geringe Cosinus-Distanz besitzen (Skalarmultiplikation von Embedding-Vektoren).
4. Die relevanten Chunks werden mit Prompt und Query an das LLM weitergeleitet und dort beantwortet
5. Das Ergebnis wird an den User zurückgegeben.

<https://aws.amazon.com/what-is/retrieval-augmented-generation/>

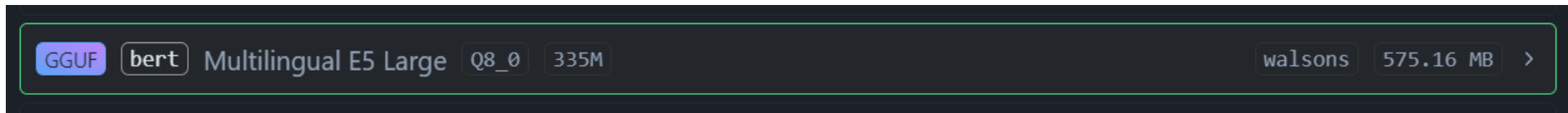


## RAG: Wie führt man Embedding durch?

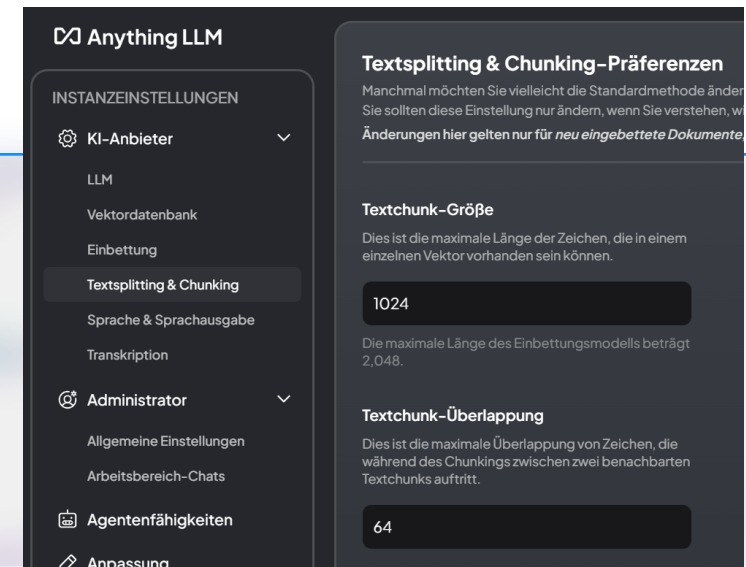
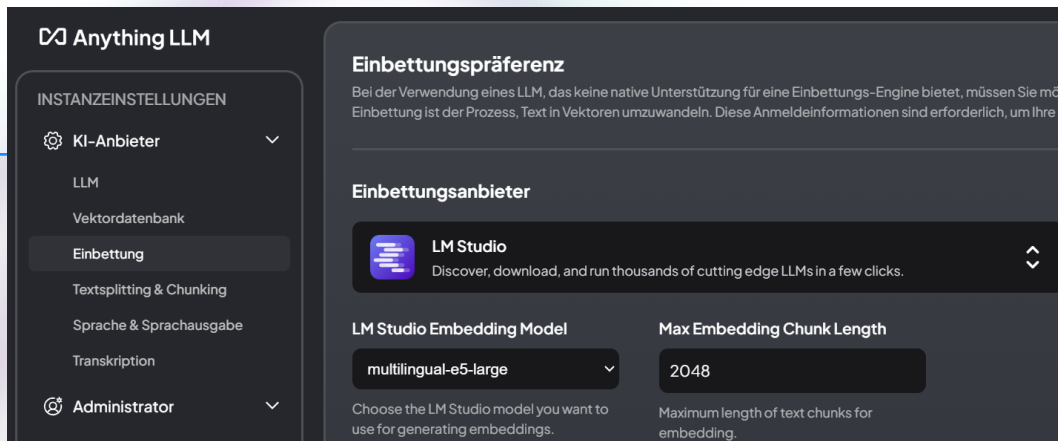
- Man benötigt ein Embedding Modell, Download innerhalb LM Studio

<https://huggingface.co/intfloat/multilingual-e5-large>

- Man lädt das Modell in LM Studio & startet den Server:



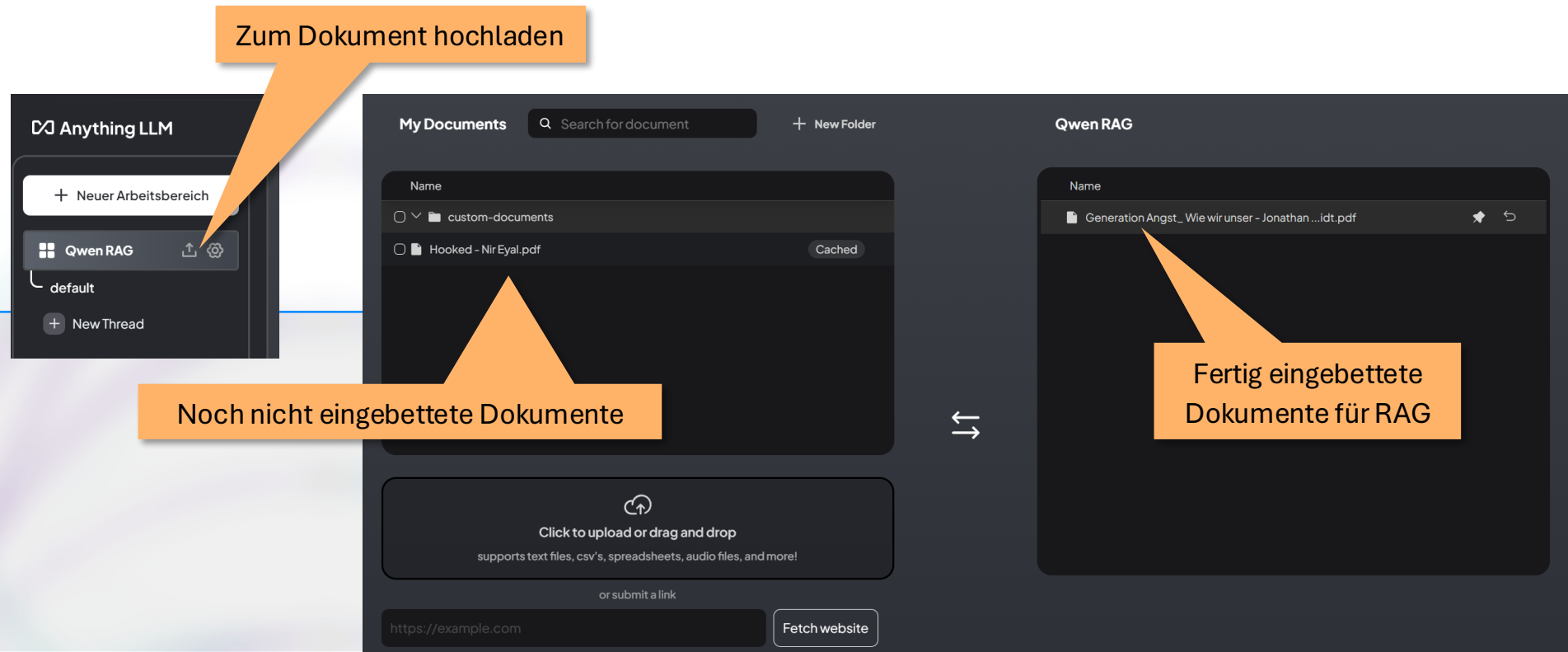
- Man konfiguriert AnythingLLM :





## RAG: Wie führt man Embedding durch?

- Im Arbeitsbereich lädt man das Dokument hoch,
- Und startet das Embedding. Das kann durchaus 10 Minuten dauern!
- Die fertigen Embeddings liegen nun in der Vektordatenbank, bereit zum Einsatz
- **Achtung! Nutzt man RAG, sind sowohl LLM als auch das Embedding-Model im Vram!**





## RAG: Wie führt man Embedding durch?

- Bevor man beginnt, stellt man die Frage „Welche Dokumente siehst du?“
- Unmittelbar auf die positive Rückmeldung des LLM kann man spezifisch fragen.

The screenshot displays the 'Anything LLM' web interface. On the left sidebar, there's a 'Neuer Arbeitsbereich' button, a 'Qwen RAG' model selector, and a 'New Thread' button. The main chat area shows a user query: 'Bitte fasse aus dem Buch "Generation Angst" von Jonathan Haidt die wichtigsten Thesen in einer Liste zusammen'. The model's response is a list of six key theses from the book 'Generation Angst' by Jonathan Haidt. At the bottom, there's a 'Send a message' input field with icons for attachments, links, mentions, and a microphone.

**Anything LLM**

+ Neuer Arbeitsbereich

Qwen RAG

default

+ New Thread

Bitte fasse aus dem Buch "Generation Angst" von Jonathan Haidt die wichtigsten Thesen in einer Liste zusammen

Natürlich, hier sind einige der wichtigsten Thesen und Argumente aus Jonathan Haidts Buch "Generation Angst", basierend auf den allgemeinen Informationen über das Thema:

01. **Technologische Entwicklung und Generationen:** Die Verbreitung von Smartphones, soziale Medien und die Selfie-Kultur haben eine neue Generation von Kindern und Jugendlichen geprägt.
02. **Psychische Gesundheit:** Der Einsatz von Technologie beeinflusst die psychische Gesundheit von Kindern und Jugendlichen negativ.
03. **Hyperviralisierung von Social-Media-Plattformen:** Die Nutzung sozialer Medien führt zu erhöhtem Vergleichsdruck, was die Selbstachtung und das Wohlbefinden beeinträchtigt.
04. **Kommunikationseinschränkungen:** Der Umgang mit virtueller Kommunikation statt persönlicher Interaktion kann die sozialen Fähigkeiten von Jugendlichen behindern.
05. **Einschränkung der Erfahrungswelt:** Die dominante Nutzung digitaler Medien verändert den Zugang zu kulturellen und gesellschaftlichen Erfahrungen, was zur Isolation und Einschränkung führt.
06. **Kritik an Elternrolle:** Es wird eine Rolle der Eltern diskutiert, wie sie ihre Kinder vor negativen Auswirkungen schützen können und welche Maßnahmen ergriffen werden sollten, um die psychische Gesundheit ihrer Kinder zu fördern.

Send a message



- 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:  $12 \text{ Kühe} * 10 \text{ Liter/Kuh} = 120 \text{ Liter}$



- 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



- MMLU (Massive Multitask Language Understanding): bewertet das Verständnis von Aufgaben in verschiedenen Bereichen, 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 im Unterricht geeignet?



- 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*
- *Agentic systems: Webscraping, Dokumentenvergleich, Beurteilung eigener Arbeiten*