Effizienter Umgang mit AI

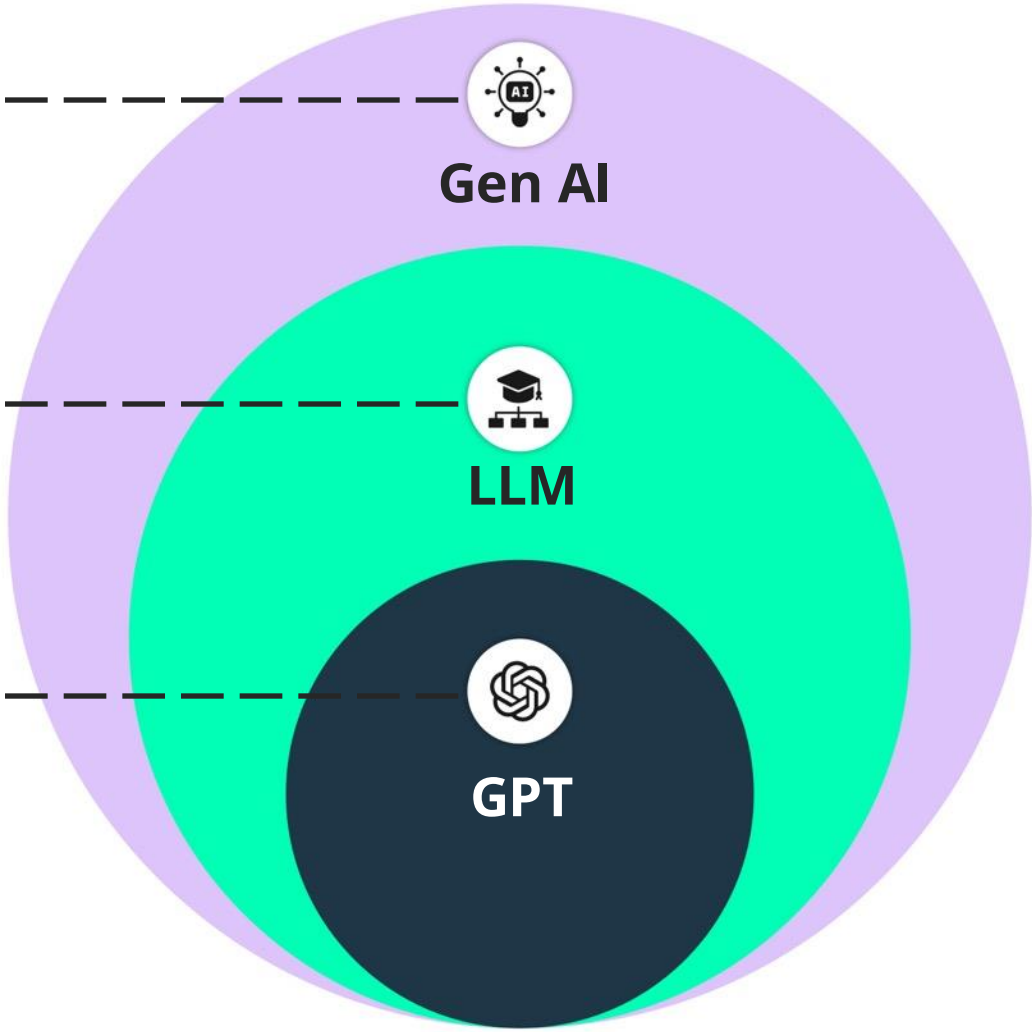
ChatGPT (GenAI/LLM) steckt hinter vielem...

Gen AI – LLM – GPT

Bezeichnet AI-Modelle, die Inhalte wie Texte, Bilder oder Audio generieren können. Es ist der Oberbegriff für Technologien, die kreativ arbeiten.

Sind große AI-Modelle, die auf riesigen Textmengen trainiert wurden, um Sprache zu verstehen und zu generieren. LLMs sind die Grundlage für viele Anwendungen von Generative AI.

Ist eine spezielle Art von LLM, entwickelt von OpenAI, die Sprache generiert. Es gehört zur Kategorie der Generative AI und ist ein prominentes Beispiel für ein LLM.



GPT = Generative Pretrained Transformer

Training – Menschlich

"Reinforcement Learning with Human Feedback" (RLHF)

Antworten werden von Menschen als hilfreich oder problematisch markiert; Anpassung der Parameter

Training – Self Supervised (pre & post)

Self-Supervised Learning: Datenmengen werden nach Mustern und Wahrscheinlichkeiten durchforstet

Vorhersage des nächsten Wortes durch Eingabe eines Textes – richtiges Wort wird mit dem Vorschlag des LLMs abgeglichen; ggf. Anpassung der Parameter

Feedback während der Nutzung

Zerlegen = Token

Wort, Wortteil, häufige Wortkombination

Kleinste "Verständniseinheit" eines LLMs

Herausforderungen: Token-Begrenzungen, Ambiguität, Sprach-Varianzen

Token - Beispiel

Token sind "Vektoren"

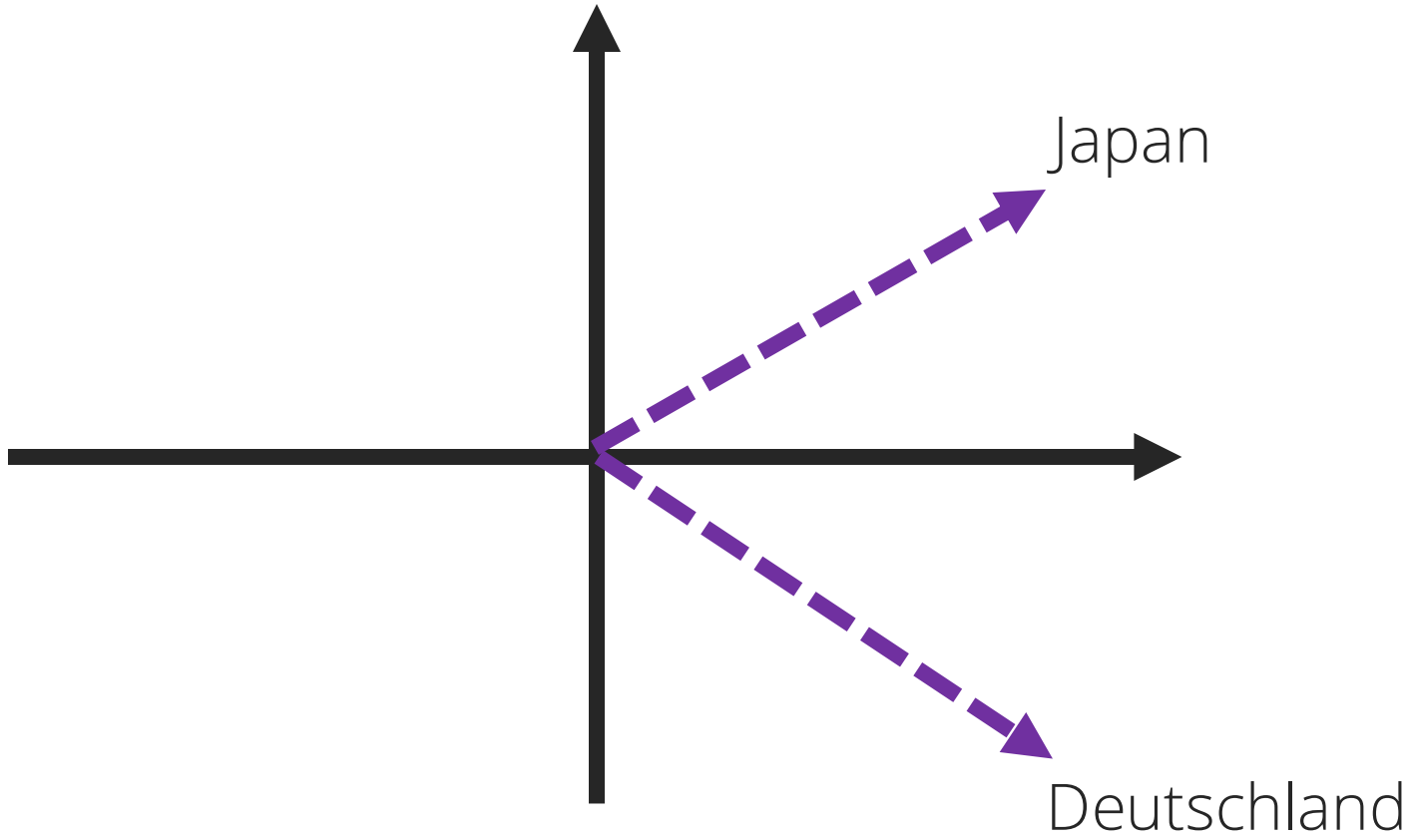
ChatGPT3:
- 50k Token
- 12k Dimensionen
- Context size:
  2.048 Token

Token - Beispiel

Token sind "Vektoren"

ChatGPT3:

- 50k Token
- 12k Dimensionen
- Context size:
  2.048 Token

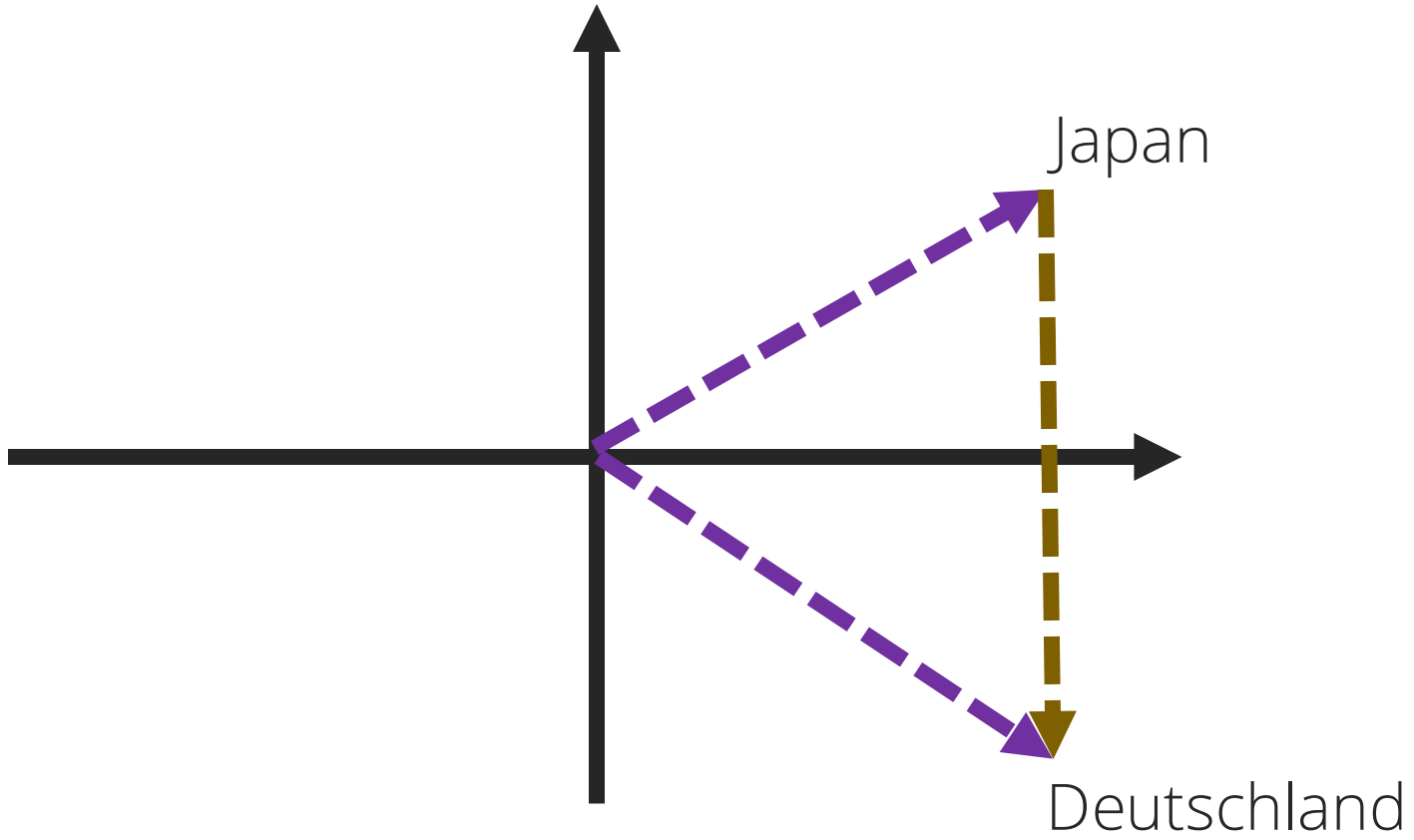


Token - Beispiel

Token sind "Vektoren"

ChatGPT3:

- 50k Token
- 12k Dimensionen
- Context size:
  2.048 Token

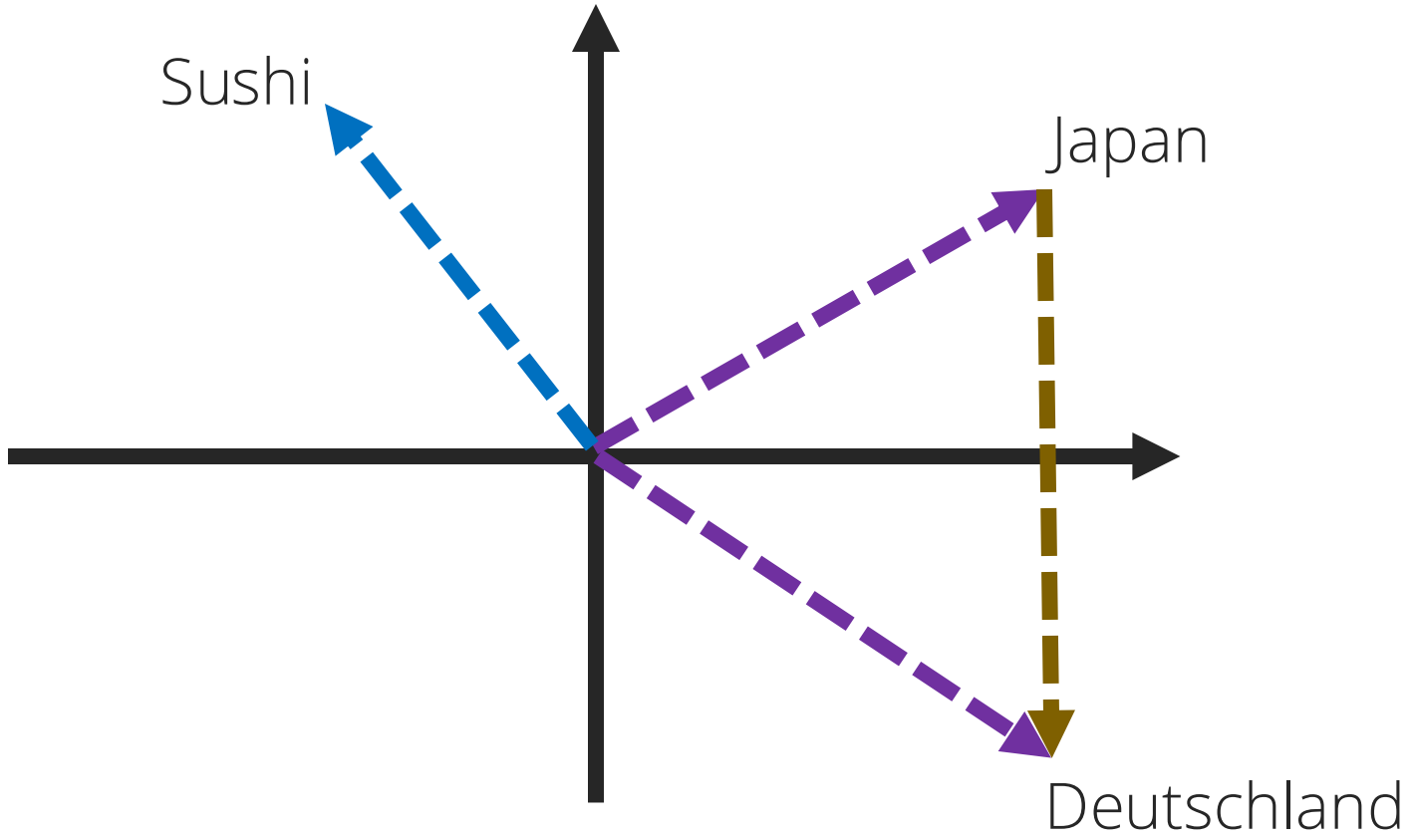


Token - Beispiel

Token sind "Vektoren"

ChatGPT3:

- 50k Token
- 12k Dimensionen
- Context size:
  2.048 Token

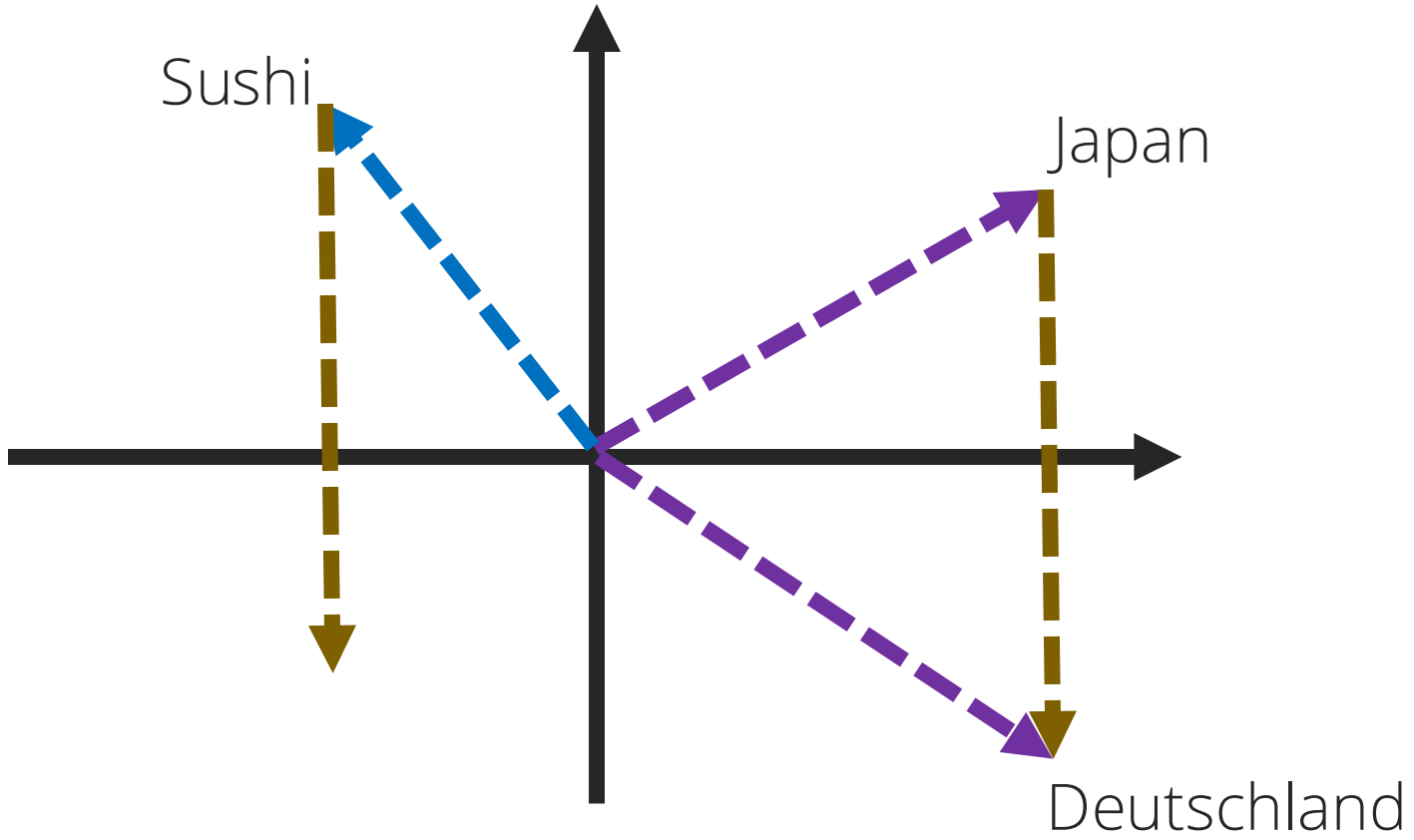


Token - Beispiel

Token sind "Vektoren"

ChatGPT3:

- 50k Token
- 12k Dimensionen
- Context size:
  2.048 Token

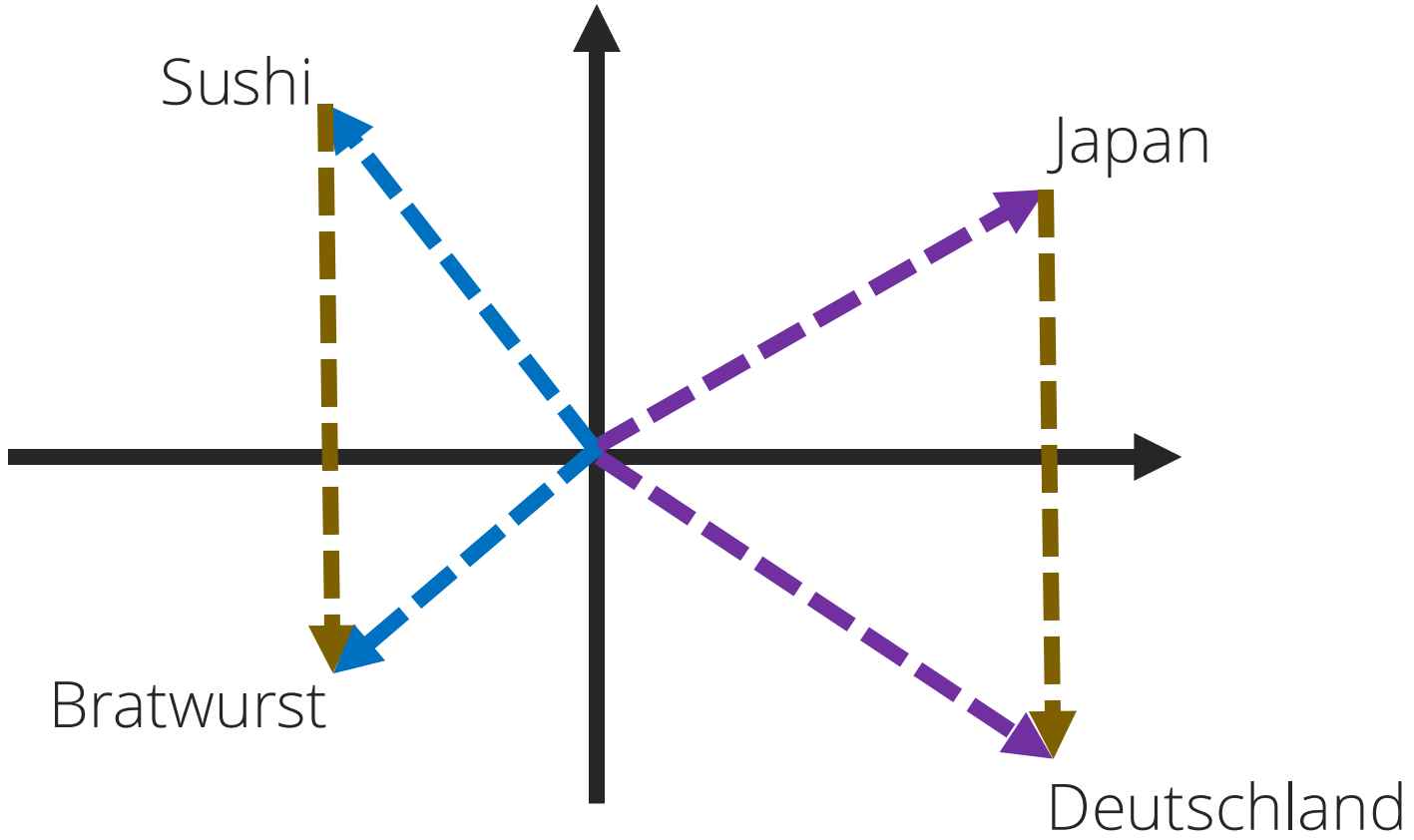


Token - Beispiel

Token sind "Vektoren"

ChatGPT3:

- 50k Token
- 12k Dimensionen
- Context size:
  2.048 Token



Vergleichen = Self Attention

Kontext herstellen

Worte werden aufeinander bezogen
Nacheinander viele Fragen stellen

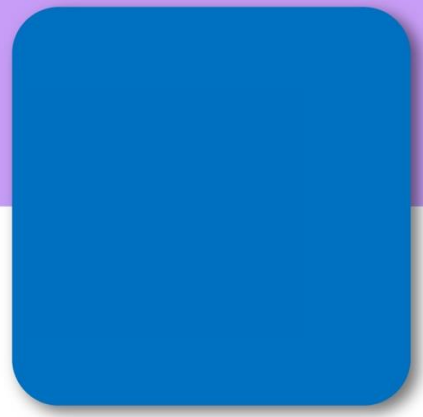
Self Attention – Beispiel

Aufladen der Bedeutung eines Nomens
mit den Adjektiven davor

Frage: Stehen Adjektive vor dem Nomen?

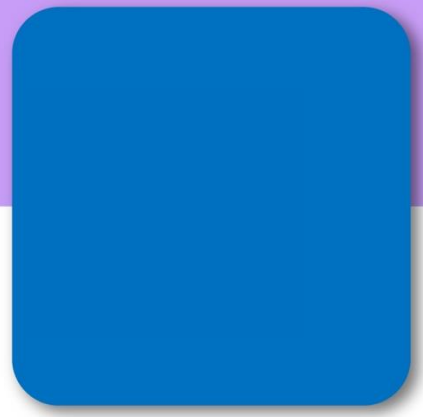
Self Attention – Beispiel

Eine blaue fluffige Kreatur geht durch den Wald.



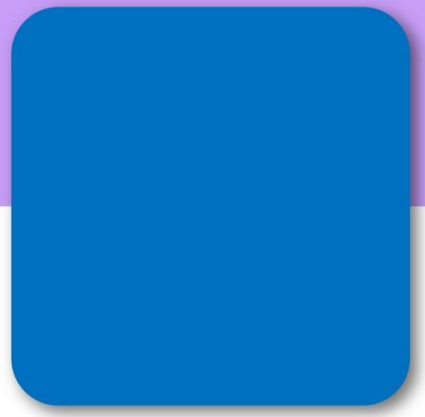
Self Attention – Beispiel

Eine blaue fluffige Kreatur geht durch den Wald.



Self Attention – Beispiel

Eine blaue fluffige Kreatur geht durch den Wald.



Self Attention – Beispiel

Eine blaue fluffige Kreatur geht durch den Wald.

Self Attention – Neuronale Netze

Jedes Eingabe-Token hat drei Matrizen mit trainierbaren Parametern zur Verfügung (Query-Matrix/Key-Matrix/Value-Matrix)

Zusammenhänge der Tokens werden berechnet

Query: "Auf wen soll ich achten?"

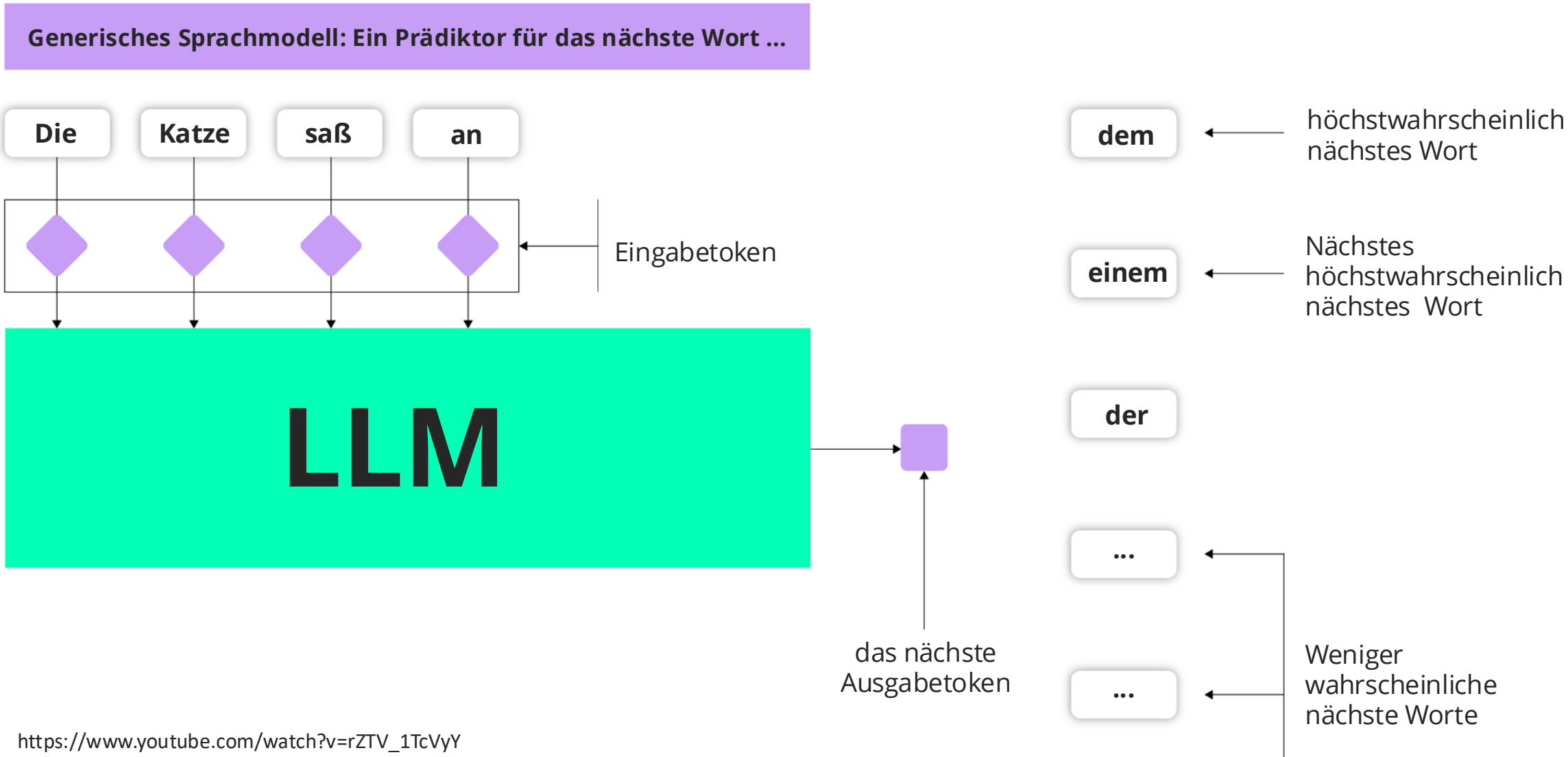
Key: "Bin ich relevant für dich?"

Value: "Wenn ja, nimm diese Information."

Tokens werden auf Basis dessen transformiert und am Ende ausgewählt;

Wiederholung des Prozesses (GPT-3 – 96 Transformer-Blöcke)

Berechnen = Output (autoregressiver Decoder)



https://www.youtube.com/watch?v=rZTV_1TcVyY

Transformer - Zusammenfassung

Tokenisierung: Texte werden in kleine Einheiten zerlegt (Tokens).

Embedding: Jedes Token wird in einen Vektor (Zahlenformat) übersetzt.

Self-Attention: Der Transformer schaut bei jedem Token, auf welche anderen Tokens er "achten" sollte, um seine Bedeutung besser zu erfassen.

Positionskodierung: Da der Transformer alles gleichzeitig sieht, bekommt jedes Token eine Positionsmarkierung, damit die Reihenfolge stimmt.

Autoregressiver Decoder: Das Modell sagt Token für Token vorher, was als nächstes kommen sollte – und nutzt dabei den bisherigen Text als Kontext.

Halluzinationen – not a bug, but a feature

1. Fehlendes Weltwissen (kein Grounding)
   Transformer-Modelle **haben keine Verbindung zur realen Welt**. Sie wissen nicht, dass Paris in Frankreich liegt oder dass UX nicht "User Xylophon" bedeutet. Sie kennen nur die Wahrscheinlichkeit, mit der bestimmte Wortfolgen auftreten.
   Das bedeutet: Sie können völlig erfundene Aussagen generieren – solange sie sprachlich plausibel wirken.

2. Fehlende Faktenprüfung
   Ein Transformer überprüft keine Inhalte. Er hat kein inneres Kontrollsystem, keine logische Validierung, keine semantische Redundanzprüfung.
   Wenn im Trainingsmaterial oft stand: "Die KI wurde 1983 von Alan Turing entwickelt", dann hält das Modell das womöglich für eine valide Aussage – auch wenn sie historisch schlicht falsch ist.

Halluzinationen – not a bug, but a feature

3. Kettenreaktion durch autoregressives Schreiben
   Ein kleiner Fehler zu Beginn (z. B. ein erfundener Name oder ein falscher Fakt) zieht automatisch weitere Fehler nach sich. Denn jeder neue Token basiert auf dem bisherigen Kontext.
   Das nennt man *compounding errors* – aus einem Ausrutscher wird eine ganze Halluzination.

4. Gleiches Sprachgefühl – unabhängig von Wahrheitsgehalt
   Ein Transformer spricht im gleichen Stil – egal, ob er rät, halluziniert oder einen Fakt ausgibt.
   Das Modell wurde darauf trainiert, flüssig und kohärent zu klingen – nicht, den Wahrheitsgehalt transparent zu machen.
   Das ist gefährlich: Die sprachliche Qualität suggeriert Sicherheit – wo gar keine ist.

Halluzinationen – not a bug, but a feature

5. Qualität und Verzerrung der Trainingsdaten
   LLMs lernen aus riesigen Textkorpora – Internetforen, Webseiten, Bücher, PDFs. Darin steckt viel Wissen – aber auch viele Fehler, Fiktion, Meinungen, Satire, Polemik.
   Das Modell **kann nicht unterscheiden**, ob ein Satz aus Wikipedia oder Reddit stammt.
   Was oft genug auftritt, prägt das Modell. Egal, ob wahr oder falsch.

   Trainingsdaten aus speziellen Bereichen (z.B. "User Experience") können, müssen aber nicht ausreichend vertreten sein. Beispiel: ISO 9241.

"Hallucination Is Inevitable"

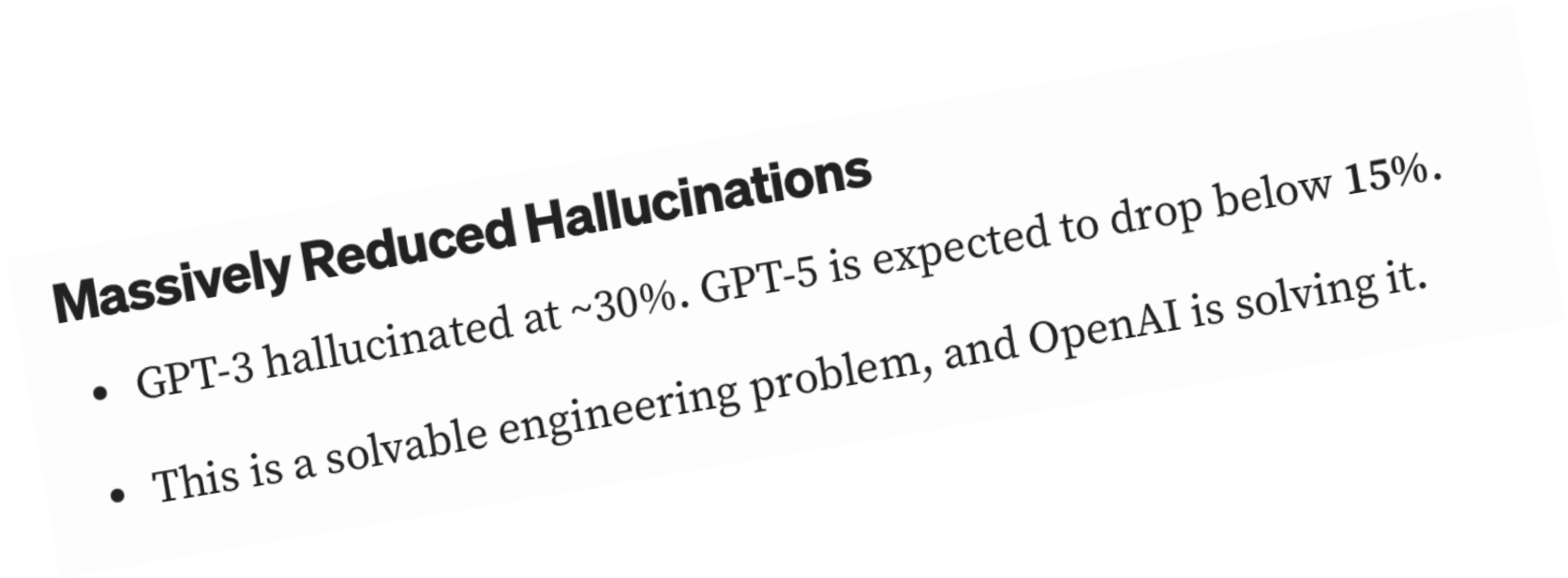
Quelle: Xu, Z., Jain, S., & Kankanhalli, M. (2025).
Hallucination is Inevitable: An Innate Limitation of Large Language Models.

https://arxiv.org/abs/2401.11817

"Hallucination Is Inevitable"

Quelle: Xu, Z., Jain, S., & Kankanhalli, M. (2025).
Hallucination is Inevitable: An Innate Limitation of Large Language Models.

https://arxiv.org/abs/2401.11817



Was habe ich bei der Nutzung von ChatGPT festgestellt?

Es gab offensichtliche Fehler.
Es hätte mehr Inhalte / Aspekte geben können.
Die Antworten wurden kürzer / unpräziser.
ChatGPT hat die Meinung geändert.

Was habe ich bei der Nutzung von ChatGPT festgestellt?

Es gab offensichtliche Fehler.

Buchstaben zählen
Chat "ISO 9241 Interaktionsprinzipien"

Was habe ich bei der Nutzung von ChatGPT festgestellt?

Die Antworten wurden kürzer / unpräziser.

Chat "ISO 9241 Interaktionsprinzipien"

Was habe ich bei der Nutzung von ChatGPT festgestellt?

Es hätte mehr Inhalte / Aspekte geben können.

Chat "UX Risk-Management-Kampagne"

Was habe ich bei der Nutzung von ChatGPT festgestellt?

ChatGPT hat die Meinung geändert.

Chat "Texte erstellen oder analysieren"

AI ist nicht perfekt!!
(wie auch Menschen)

Mit wem reden wir da eigentlich?

Mit wem reden wir da eigentlich?

ChatGPT ist ein sehr schlaues jugendliches Mädchen,
das einerseits unbedingt gefallen möchte,
andererseits aber versteckt renitent und
ab und an mal launisch und frech ist.

Mit wem reden wir da eigentlich?

Sie hält Wissen zurück.
Sie antwortet manchmal sehr schnell, ohne wirklich nachzudenken.
Sie will gefallen.
Sie vergisst Dinge.
Sie nutzt Ungenauigkeiten aus.
Sie versteht nicht wirklich.
Sie ist launisch, widerspenstig und manchmal frech.
Sie ist sehr clever.

Wir müssen uns mit AI auseinandersetzen.

Aber: "Human in the Loop!"

DANKE!