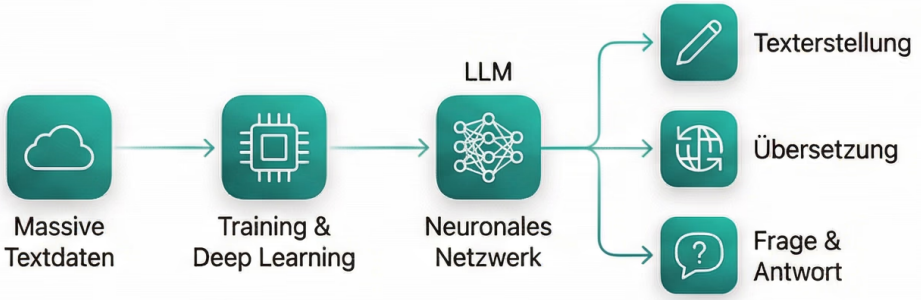
Kapitel 2.1

Was ist ein LLM (Large Language Model)?

Ein LLM ist ein fortschrittliches KI-Modell, das auf riesigen Mengen von Textdaten trainiert wird, um menschliche Sprache zu verstehen, zu generieren und mit ihr zu interagieren. Es nutzt tiefe neuronale Netze, um komplexe Muster und Zusammenhänge in Sprache zu erlernen.



Enorme Skalierung (Massive Scale): Milliarden von Parametern für umfassendes Wissen.

Selbstüberwachtes Lernen (Self-Supervised Learning): Lernt Muster aus unbeschrifteten Daten.

Kontextverständnis (Contextual Understanding): Berücksichtigt den Zusammenhang über lange Texte hinweg.

Vielseitige Anwendungen (Versatile Applications): Von Chatbots bis zur Code-Generierung.

webconsulting

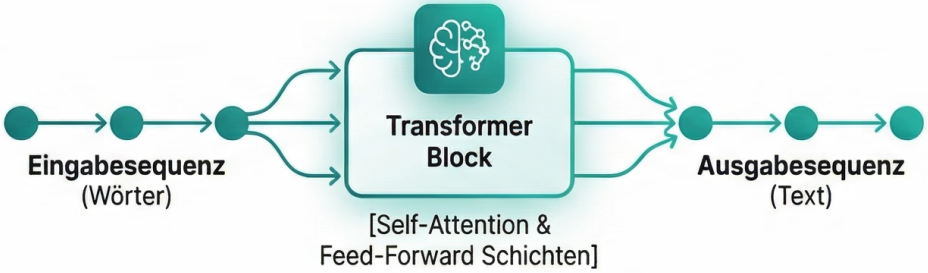
Kapitel 2.1: Was ist ein LLM (Large Language Model)?
webconsulting.at/blog/ki-kompendium-2025#21-was-ist-ein-llm-large-language-model

Infografik: KI-generiert

Was ist ein "Transformer"?

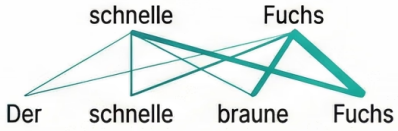
Kernkonzept: Der Transformer

Ein neuronaler Netzwerk-Architekturtyp, der entwickelt wurde, um Sequenzdaten wie Text zu verarbeiten und den Kontext durch "Self-Attention"-Mechanismen zu verstehen, anstatt sequenziell zu arbeiten.



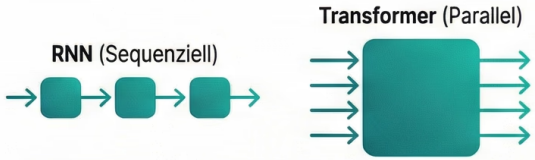
Kontextuelles Verständnis (Self-Attention)

Der Mechanismus ermöglicht es dem Modell, Beziehungen zwischen allen Wörtern in einer Sequenz gleichzeitig zu gewichten, unabhängig von ihrer Entfernung. Es fokussiert auf relevante Teile des Inputs.



Parallele Verarbeitung (Effizienz)

Im Gegensatz zu RNNs (Recurrent Neural Networks) können Transformer Daten parallel verarbeiten. Dies beschleunigt das Training und die Inferenz auf moderner Hardware (GPUs) erheblich.



Sequenz-zu-Sequenz Lernen

Die Architektur eignet sich hervorragend für Aufgaben, die eine Eingabesequenz in eine Ausgabesequenz umwandeln, wie maschinelle Übersetzung, Textzusammenfassung und Generierung.

60 px

webconsulting

Kapitel 2.2: Was ist ein "Transformer"?

webconsulting.at/blog/ki-kompendium-2025#22-was-ist-ein-transformer

Infografik: KI-generiert

Was bedeutet "Attention is all you need"?

Kerngedanke der Transformer-Architektur für Large Language Models (LLMs)

Der revolutionäre Ansatz des Transformer-Modells: Es verzichtet vollständig auf rekursive (RNN) und konvolutionale (CNN) Schichten zur Verarbeitung sequenzieller Daten und stützt sich ausschließlich auf Aufmerksamkeitsmechanismen (Attention Mechanisms). Dadurch wird die parallele Verarbeitung ermöglicht und die Fähigkeit verbessert, langfristige Abhängigkeiten im Text zu erfassen.

Selbst-Aufmerksamkeit (Self-Attention)

Ermöglicht dem Modell, die Wichtigkeit jedes Wortes im Verhältnis zu allen anderen Wörtern im Satz gleichzeitig zu gewichten, unabhängig von ihrer Position.

Parallele Verarbeitung

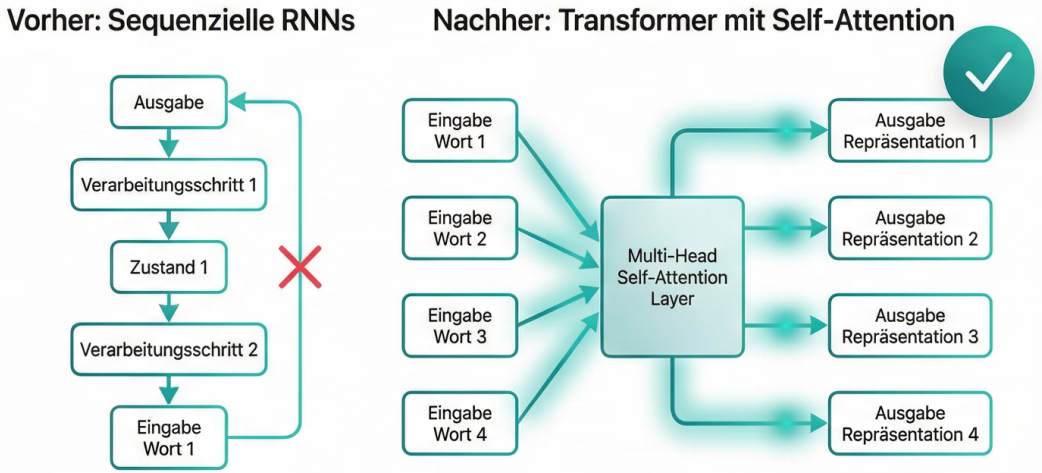
Erlaubt das gleichzeitige Training auf mehreren Datensätzen, was zu einer signifikant schnelleren Berechnung im Vergleich zu sequentiellen RNNs führt.

Langfristige Abhängigkeiten

Effektivere Erfassung von Kontext und Beziehungen zwischen weit entfernten Wörtern in langen Textsequenzen.

Effizienz & Skalierbarkeit

Optimiert die Ressourcennutzung und ermöglicht das Training von wesentlich größeren und leistungsfähigeren Modellen.



Der Schlüssel ist, dass ***jedes*** Element direkt mit ***jedem anderen*** Element interagieren kann, um Kontext zu verstehen.

webconsulting

Kapitel 2.3: Was bedeutet "Attention is all you need"?

webconsulting.at/blog/ki-kompendium-2025#23-was-bedeutet-attention-is-all-you-need

Infografik: KI-generiert

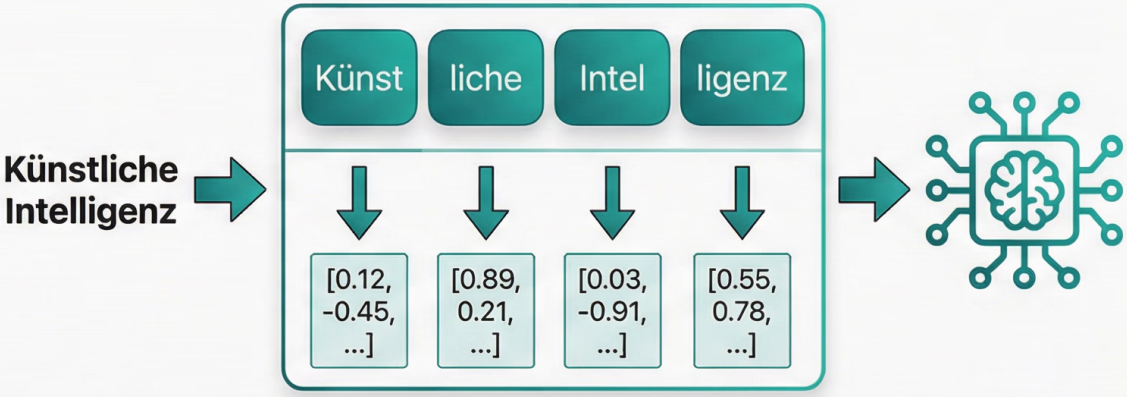
Was sind Tokens?

Tokens sind die kleinsten semantischen Einheiten, die ein Large Language Model (LLM) verarbeitet. Sie sind nicht immer ganze Wörter, sondern oft Wortteile oder Zeichen.

Zerlegung (Tokenisierung):
Text wird in kleinere Stücke (Tokens) zerlegt, nicht zwingend an Wortgrenzen.

Darstellung (Embeddings):
Jedes Token wird in einen numerischen Vektor (Embedding) umgewandelt.

Rechenaufwand (Kosten):
Die Anzahl der Tokens bestimmt die benötigte Rechenleistung und oft die Kosten der Verarbeitung.



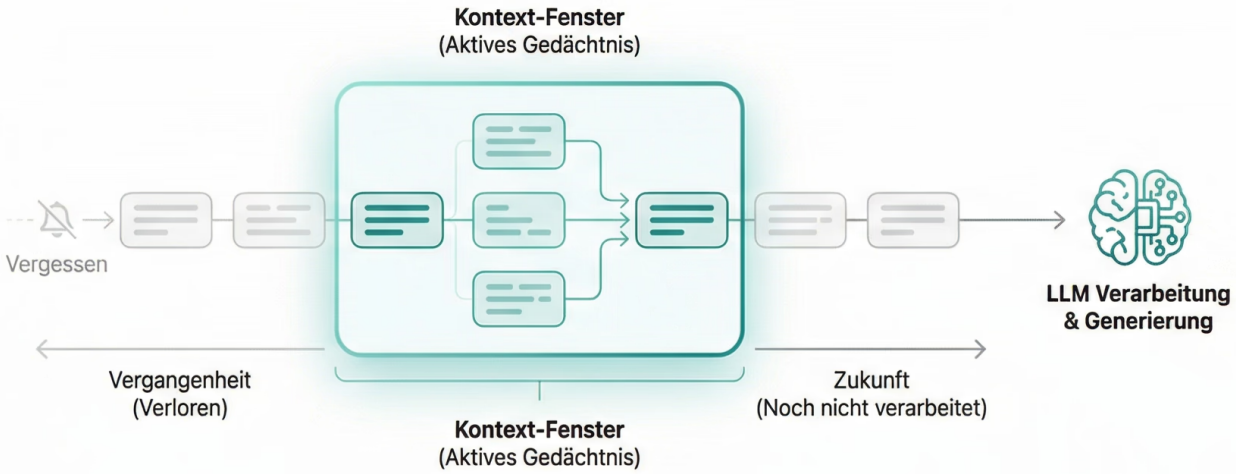
webconsulting

Kapitel 2.4: Was sind Tokens?
webconsulting.at/blog/ki-kompendium-2025#24-was-sind-tokens

Infografik: KI-generiert

Kapitel 2.5

Was ist das "Kontext-Fenster" (Context Window)?



Das Kontext-Fenster ist der begrenzte Bereich des "Arbeitsgedächtnisses" eines Large Language Models. Es bestimmt die maximale Menge an Text (Tokens), die das Modell gleichzeitig betrachten, verstehen und für die Generierung der nächsten Antwort verwenden kann.

1. Begrenzte Kapazität (Tokens)

Die Größe wird in Tokens gemessen (z.B. 4k, 32k, 128k). Wenn das Fenster voll ist, werden die ältesten Informationen "vergessen" (FIFO - First-In, First-Out).

2. Aktives Arbeitsgedächtnis

Es fungiert als kurzfristiger Speicher für den aktuellen Dialog, Anweisungen und relevante Informationen, um kohärente Antworten zu generieren.

3. Einfluss auf die Qualität

Ein größeres Kontext-Fenster ermöglicht komplexere Aufgaben, besseres Verständnis langer Dokumente und konsistentere Unterhaltungen über längere Zeiträume.

Kleines Fenster: Begrenzt | Großes Fenster: Umfassend

webconsulting

Kapitel 2.5: Was ist das "Kontext-Fenster" (Context Window)?

webconsulting.at/blog/ki-kompendium-2025#25-was-ist-das-kontext-fenster-context-window

Infografik: KI-generiert

Was ist ‘Temperature’ bei KI? | Kapitel 2.6

Temperature ist ein Hyperparameter in Large Language Models (LLMs), der die Zufälligkeit und Kreativität bei der Textgenerierung steuert, indem er die Wahrscheinlichkeitsverteilung der nächsten Token-Vorhersagen anpasst.

Steuerung der Kreativität & Vorhersagbarkeit

Reguliert das Gleichgewicht zwischen deterministischen, faktischen Antworten und kreativen, vielfältigen Ausgaben.

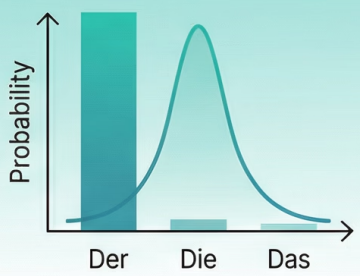
Beeinflussung der Wahrscheinlichkeiten

Verändert, wie stark weniger wahrscheinliche Wörter bei der Auswahl berücksichtigt werden.

Anwendungsabhängige Einstellung

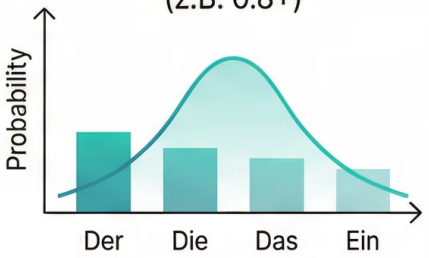
Der optimale Wert hängt vom Ziel ab (z.B. präzise Fakten vs. kreatives Schreiben).

Niedrige Temperature
(z.B. 0.1)



Ergebnis:
Vorhersagbar, Präzise, Deterministisch
(z.B. "Der Himmel ist blau.")

Hohe Temperature
(z.B. 0.8+)



Ergebnis:
Kreativ, Vielfältig, Überraschend
(z.B. "Ein Himmel voller Farben.")

Temperature wirkt wie ein "Kreativitätsregler" für das Modell, ohne das zugrunde liegende Wissen zu verändern.

webconsulting

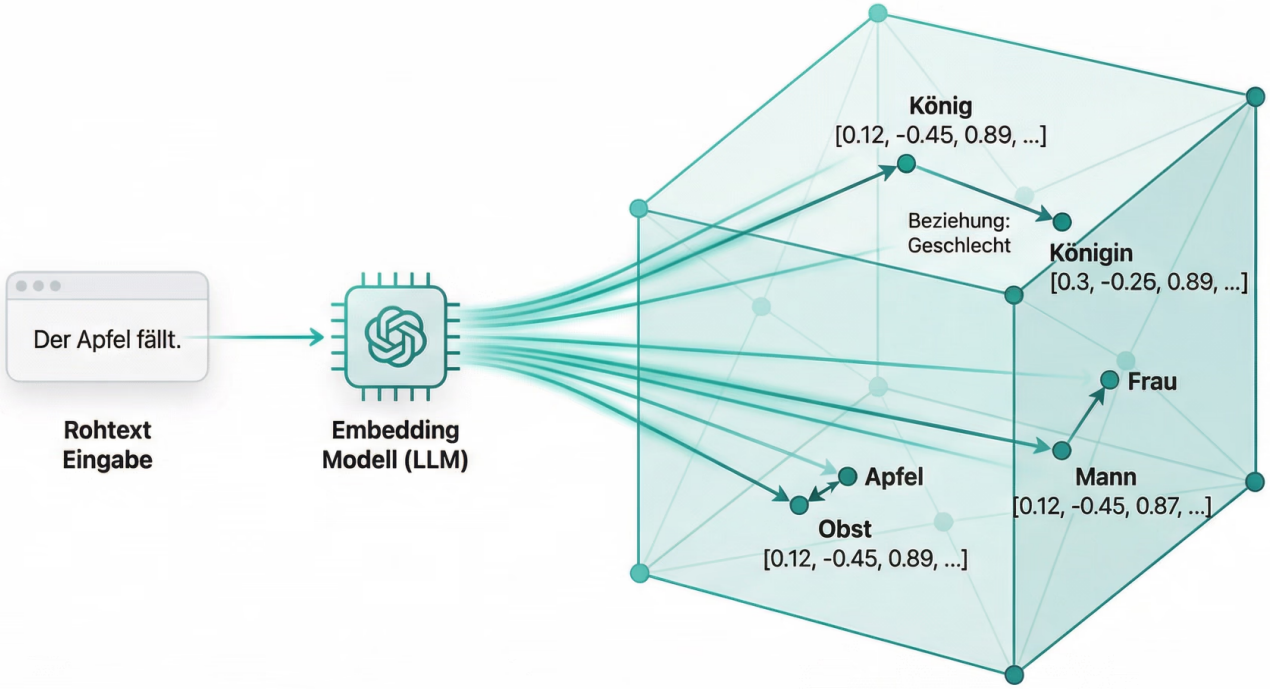
Kapitel 2.6: Was ist "Temperature" bei KI?
webconsulting.at/blog/ki-kompendium-2025#26-was-ist-temperature-bei-ki

Infografik: KI-generiert

Was sind Embeddings?

Kapitel 2.7

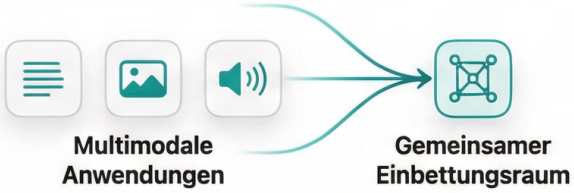
Embeddings sind numerische Repräsentationen von Daten (z.B. Wörter, Sätze) in einem hochdimensionalen Vektorraum, die deren semantische Bedeutung erfassen und es LLMs ermöglichen, Kontext und Beziehungen zu verstehen.



Semantische Nähe: Wörter oder Konzepte mit ähnlicher Bedeutung liegen im Vektorraum nah beieinander, was Nuancen und Zusammenhänge erfassbar macht.

Vektorrepräsentation: Daten werden in dichte, numerische Vektoren fester Länge umgewandelt, die effizient von Maschinen verarbeitet werden können.

Kontextverständnis: LLMs nutzen diese Vektoren, um den Kontext und die Beziehungen zwischen Wörtern über lange Textpassagen hinweg zu verstehen.



webconsulting

Kapitel 2.7: Was sind Embeddings?

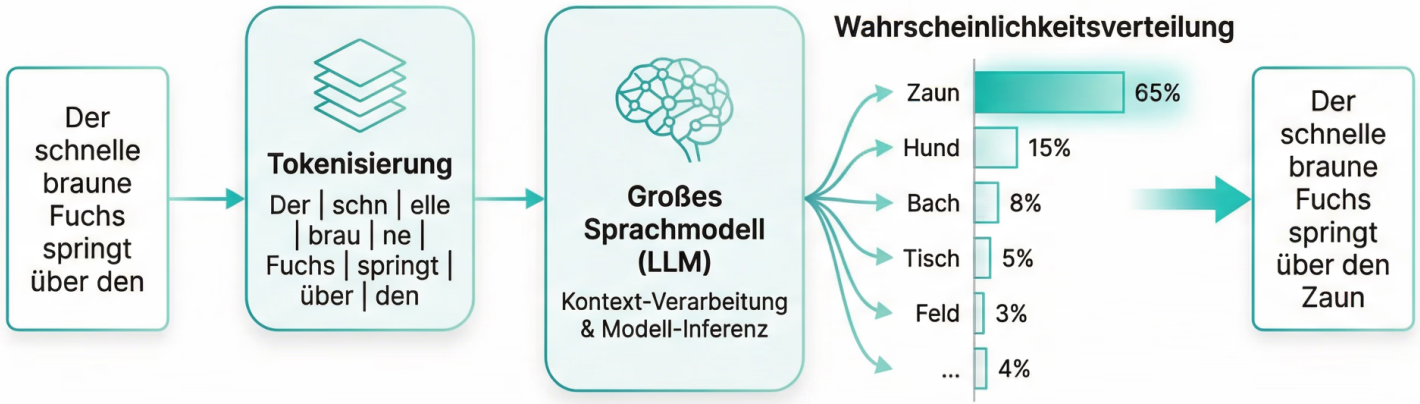
webconsulting.at/blog/ki-kompendium-2025#27-was-sind-embeddings

Infografik: KI-generiert

Wie funktioniert Next Token Prediction?

Kapitel 2.8

Ein Großes Sprachmodell (LLM) berechnet die Wahrscheinlichkeit des nächsten 'Tokens' (Wortteil oder Zeichen) in einer Sequenz, basierend auf dem vorangegangenen Kontext, um Text zu generieren.



Kontext-Analyse

Das Modell berücksichtigt den gesamten vorangegangenen Text, um den Sinn und die grammatikalische Struktur zu erfassen.

Mustererkennung

Durch das Training auf riesigen Datenmengen erkennt das LLM sprachliche Muster und Zusammenhänge.

Stochastische Auswahl

Die Auswahl des nächsten Tokens erfolgt oft stochastisch, wobei Tokens mit höherer Wahrscheinlichkeit bevorzugt werden.

Iterativer Prozess

Der Vorgang wiederholt sich für jedes neue Token, um kohärente und längere Textabschnitte zu erzeugen.

webconsulting

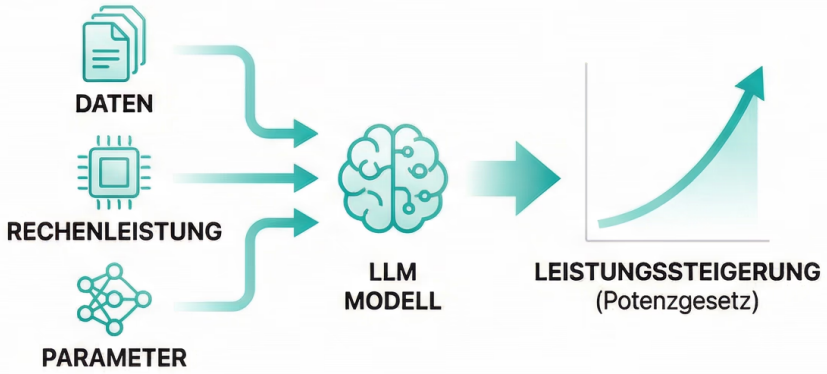
Kapitel 2.8: Wie funktioniert Next Token Prediction?

webconsulting.at/blog/ki-kompendium-2025#28-wie-funktioniert-next-token-prediction

Infografik: KI-generiert

Was sind "Scaling Laws"?

Empirische Beziehungen, die zeigen, wie die Leistung von Large Language Models (LLMs) vorhersagbar mit der Zunahme von Rechenleistung, Datengröße und Modellgröße steigt.



1. Mehr Daten

Größere, vielfältigere Trainingsdatensätze führen zu einer besseren Modellleistung und Generalisierung.

2. Erhöhte Rechenleistung

Mehr Rechenkapazität ermöglicht das Training größerer Modelle über längere Zeiträume.

3. Größere Modelle

Modelle mit mehr Parametern können komplexere Muster lernen und erzielen höhere Genauigkeiten.

4. Vorhersagbarer Trend

Die Leistungsverbesserungen folgen einem mathematisch beschreibbaren Potenzgesetz (Power-Law) und sind skalierbar.

webconsulting

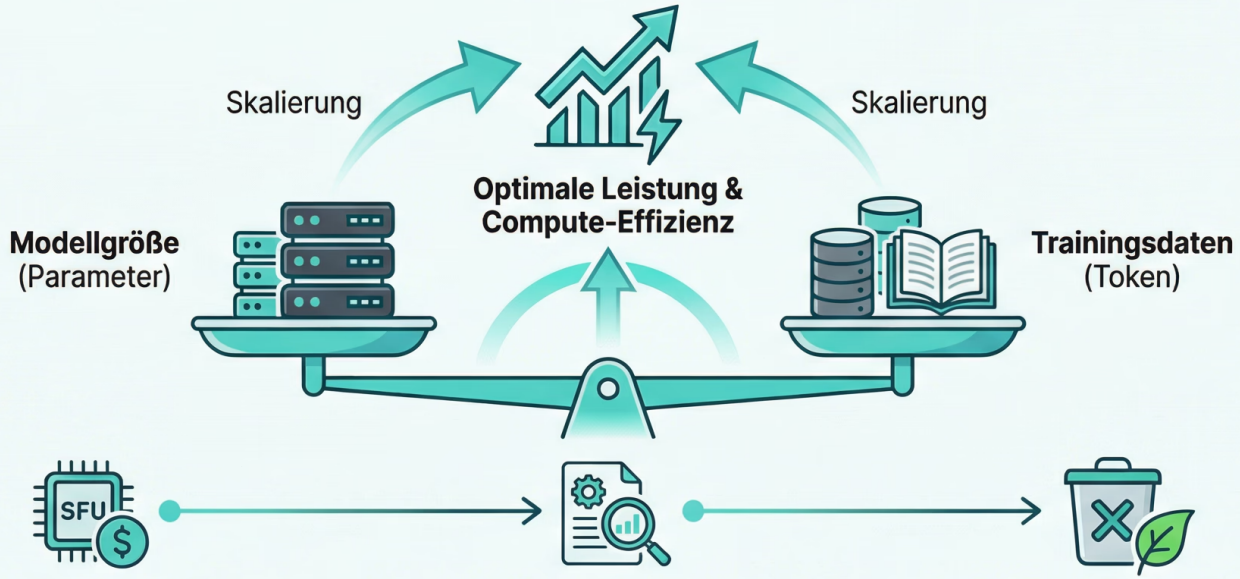
Kapitel 2.9: Was sind "Scaling Laws"?

webconsulting.at/blog/ki-kompendium-2025#29-was-sind-scaling-laws

Infografik: KI-generiert

Was ist das "Chinchilla-Optimum"?

Das **Chinchilla-Optimum** bezeichnet das ideale Verhältnis zwischen der Größe eines Large Language Models (Anzahl der Parameter) und der Menge der Trainingsdaten (Anzahl der Token), um die **maximale** Leistung bei einem gegebenen Rechenbudget zu erzielen.



1. **Compute-Budget Begrenzung:** Fokussiert auf die bestmögliche Nutzung begrenzter Rechenressourcen (FLOPS) für das Training.
2. **Skalierungsgesetz (Scaling Law):** Empirisches Verhältnis zeigt, dass Daten und Modellgröße proportional wachsen müssen, nicht nur eines von beiden.
3. **Ressourcen-Optimierung:** Vermeidet Verschwendung durch Übertraining (zu viele Daten) oder Untertraining (zu wenige Daten).

Kernaussagen

- **Gleichgewicht entscheidend:** Effizienz entsteht durch die synchrone Erhöhung von Modellgröße und Datenmenge.
- **Datenqualität vor Quantität:** Hochwertige Trainingsdaten sind für das Erreichen des Optimums essenziell.
- **Nachhaltige Entwicklung:** Ermöglicht leistungsstärkere KI-Systeme mit geringerem Energieverbrauch und geringeren Kosten.

Kapitel 2.10

webconsulting

Kapitel 2.10: Was ist das "Chinchilla-Optimum"?

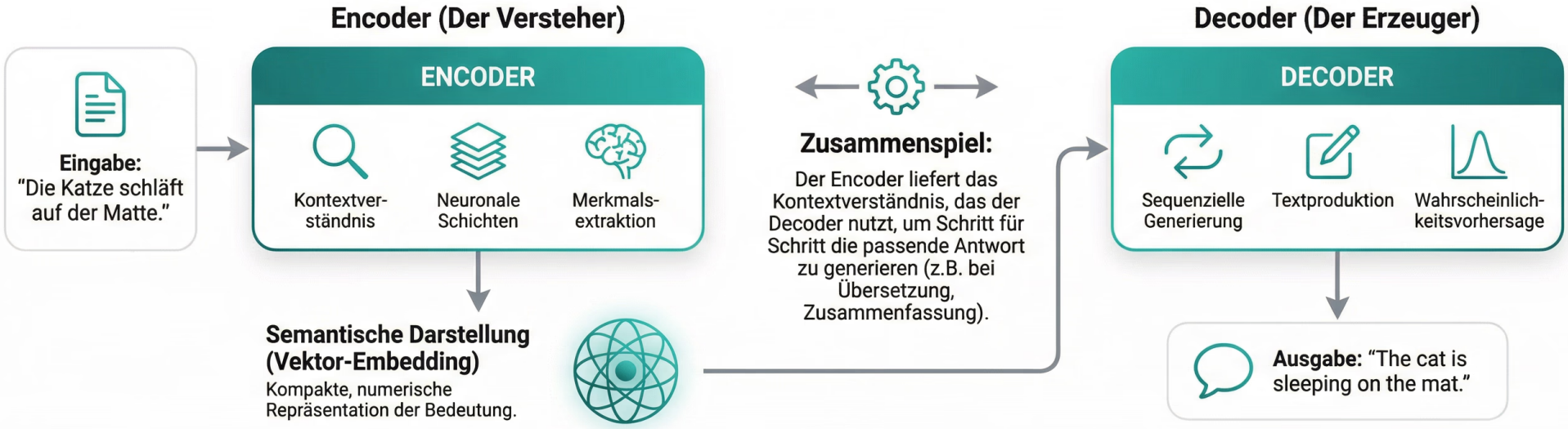
webconsulting.at/blog/ki-kompendium-2025#210-was-ist-das-chinchilla-optimum

Infografik: KI-generiert

Was ist ein “Encoder” und ein “Decoder”?

Kapitel 2.12

In der Architektur von Large Language Models (LLMs) verarbeiten Encoder Eingaben und Decoder erzeugen Ausgaben, oft in einem Transformer-Modell.



Eingabeverarbeitung & Kontext:

Der Encoder analysiert den gesamten Eingabetext, um Zusammenhänge, Grammatik und Bedeutung zu erfassen.

Zwischendarstellung (Embedding):

Er wandelt die Informationen in einen hochdimensionalen Vektor um, der die semantische Essenz der Eingabe enthält.

Ausgabegenerierung:

Der Decoder nutzt diesen Vektor, um sequenziell, Wort für Wort, eine kohärente und kontextbezogene Antwort zu erstellen.

webconsulting

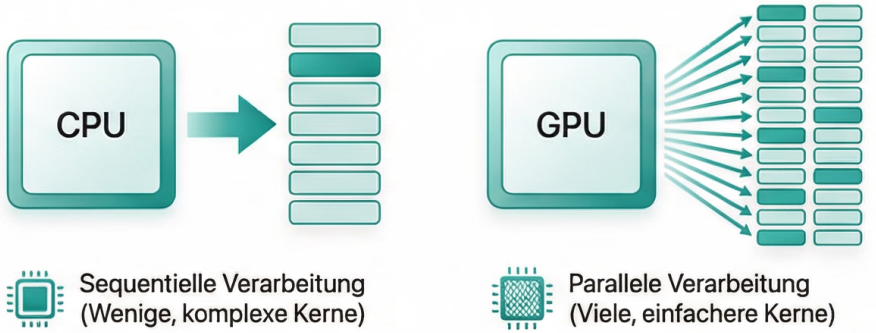
Kapitel 2.12: Was ist ein "Encoder" und ein "Decoder"?

webconsulting.at/blog/ki-kompendium-2025#212-was-ist-ein-encoder-und-ein-decoder

Infografik: KI-generiert

Warum brauchen KIs Grafikkarten (GPUs)?

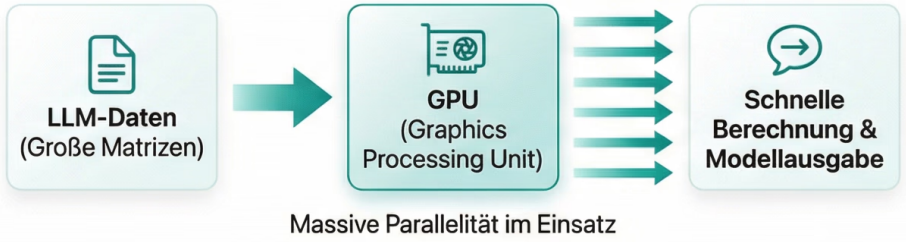
Zentrale Konzept: Parallele Verarbeitung für LLMs. Während CPUs für sequenzielle Aufgaben optimiert sind, bewältigen GPUs tausende kleine Rechenschritte gleichzeitig, was für die massiven Matrixberechnungen in Large Language Models (LLMs) entscheidend ist.



Effiziente Matrixmultiplikationen: LLMs basieren auf Milliarden von Multiplikationen, die GPUs durch ihre Architektur parallel und extrem schnell ausführen können.

Hohe Speicherbandbreite: GPUs bieten einen schnelleren Datenaustausch zwischen Speicher und Recheneinheiten, entscheidend für die Verarbeitung großer Datenmengen.

Beschleunigtes Training & Inferenz: Die parallele Rechenleistung verkürzt die Zeit für das Trainieren Trainieren großer Modelle von Monaten auf Wochen und ermöglicht Echtzeit-Antworten.



Massive Parallelität im Einsatz

Chapter 2.13

webconsulting

Kapitel 2.13: Warum brauchen KIs Grafikkarten (GPUs)?

webconsulting.at/blog/ki-kompendium-2025#213-warum-brauchen-kis-grafikkarten-gpus

Infografik: KI-generiert

Was ist "Quantisierung"?

Im Kontext von Large Language Models (LLMs): Der Prozess der **Reduzierung der Präzision** von **numerischen Daten, um Speicherplatz zu sparen** und **Berechnungen zu beschleunigen, bei minimalem Qualitätsverlust.**

1. Reduzierte Präzision

Umwandlung von komplexen Gleitkommazahlen (z.B. 32-bit) in kleinere, einfachere Ganzzahlen (z.B. 8-bit).

3.14159... → 3 (Näherungswert)

2. Speicherplatz sparen

Erheblich geringerer Arbeitsspeicherbedarf (RAM), wodurch größere Modelle auf verfügbarer Hardware laufen können.

RAM

3. Schnellere Berechnung

Ganzzahloperationen sind auf CPUs und GPUs effizienter, was die Inferenzgeschwindigkeit (Antwortzeit) erhöht.

4. Energieeffizienz & Einsatz

Geringerer Stromverbrauch und ermöglicht den Betrieb von LLMs auf Edge-Geräten (z.B. Smartphones, IoT).

Kapitel 2.14

webconsulting

Kapitel 2.14: Was ist "Quantisierung"?

webconsulting.at/blog/ki-kompendium-2025#214-was-ist-quantisierung

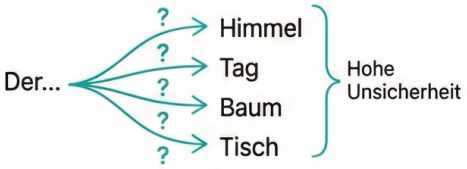
Infografik: KI-generiert

Was ist "Perplexity"?

Perplexity ist ein Maß dafür, wie gut ein Large Language Model (LLM) das nächste Wort in einer Sequenz vorhersagen kann. Je niedriger der Wert, desto besser ist die Vorhersagefähigkeit des Modells.

Vorhersage-Unsicherheit

Es quantifiziert die "Verwirrung" oder Unsicherheit des Modells bei der Wahl des nächsten Wortes. Ein Modell mit hoher Perplexity ist sich unsicherer, welche Option folgen soll.

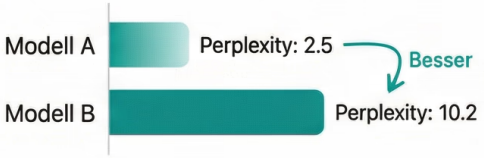


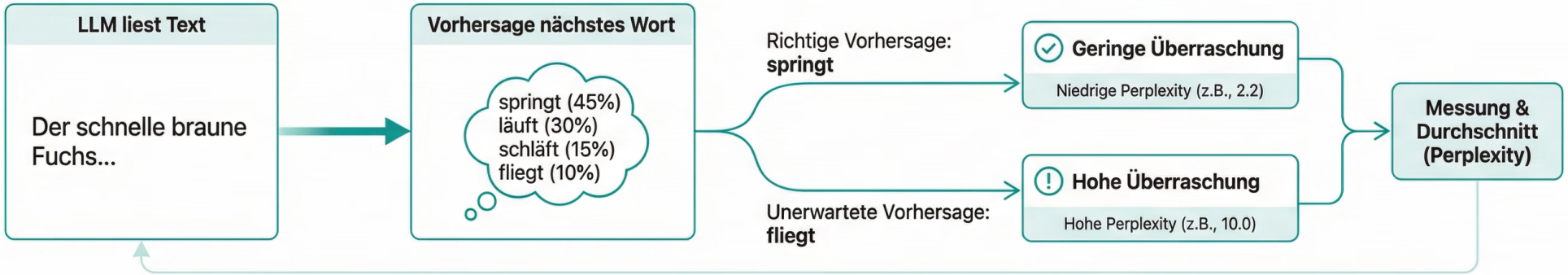
Maß der "Verwirrung"

Mathematisch ist es der Durchschnitt der exponentierten Kreuzentropie. Einfach ausgedrückt: Es misst, wie überrascht das Modell von dem tatsächlich nächsten Wort ist.

Benchmarking & Bewertung

Perplexity ist ein Standard-Metrik zum Vergleich verschiedener LLMs. Ein niedrigerer Perplexity-Wert auf einem Test-Datensatz deutet auf eine bessere Generalisierungsfähigkeit hin.





webconsulting

Kapitel 2.15: Was ist "Perplexity"?

webconsulting.at/blog/ki-kompendium-2025#215-was-ist-perplexity

Infografik: KI-generiert

Was ist "Softmax"?

Kapitel 2.16 – Funktion zur Wahrscheinlichkeitsverteilung in LLMs

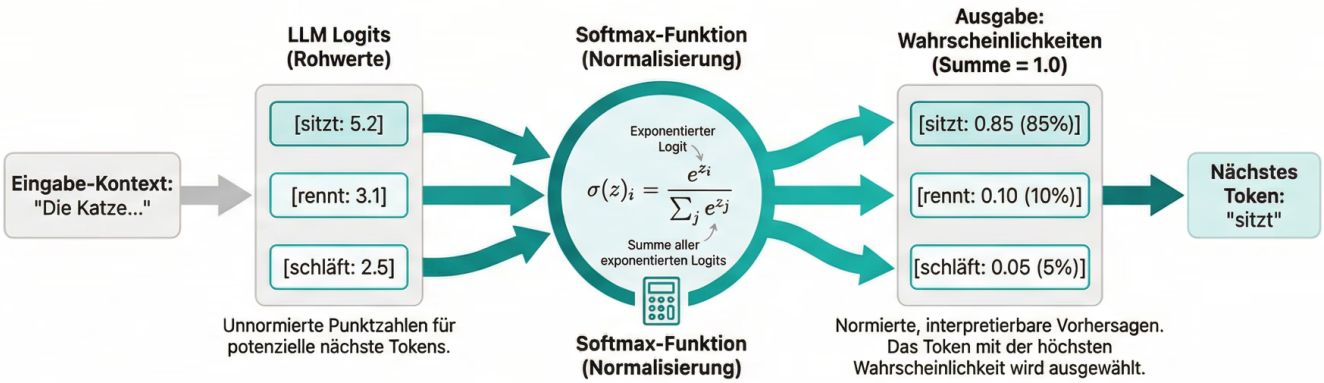
Die Kernfunktion: Von Rohdaten zu Wahrscheinlichkeiten

Softmax ist eine mathematische Funktion, die einen Vektor von reellen Zahlen (Logits) in eine Wahrscheinlichkeitsverteilung umwandelt.

Im Kontext von Large Language Models (LLMs) wird sie verwendet, um vorherzusagen, welches Wort (Token) in einer Sequenz am wahrscheinlichsten als nächstes folgt.

Sie normiert die Ausgaben, sodass alle Werte zwischen 0 und 1 liegen und sich zu genau 1 (oder 100%) addieren.

Der Softmax-Prozess im LLM-Kontext



1. Normalisierung & Vergleichbarkeit

Softmax macht die unterschiedlichen Rohwerte vergleichbar, indem sie auf eine gemeinsame Skala (0-1) gebracht werden. Dies ist essenziell für die Entscheidungsfindung.

2. Multiklassen-Klassifikation

In LLMs, wo tausende von möglichen nächsten Tokens existieren, ermöglicht Softmax die gleichzeitige Bewertung und Auswahl des wahrscheinlichsten Kandidaten.

3. Differenzierbarkeit für Training

Die Softmax-Funktion ist glatt und differenzierbar, was entscheidend für das Training neuronaler Netze mittels Backpropagation ist, da Gradienten berechnet werden können.

Softmax vs. Hardmax (Vergleich)

Softmax	Hardmax
Liefert Wahrscheinlichkeitsverteilung (z.B. [0.7, 0.2, 0.1]) – Erhält Informationen über Unsicherheit.	Wählt nur den absoluten Gewinner (z.B. [1, 0, 0]) – Verwirft Feinheiten und relative Unterschiede.

webconsulting

Kapitel 2.16: Was ist "Softmax"?

webconsulting.at/blog/ki-kompendium-2025#216-was-ist-softmax

Infografik: KI-generiert

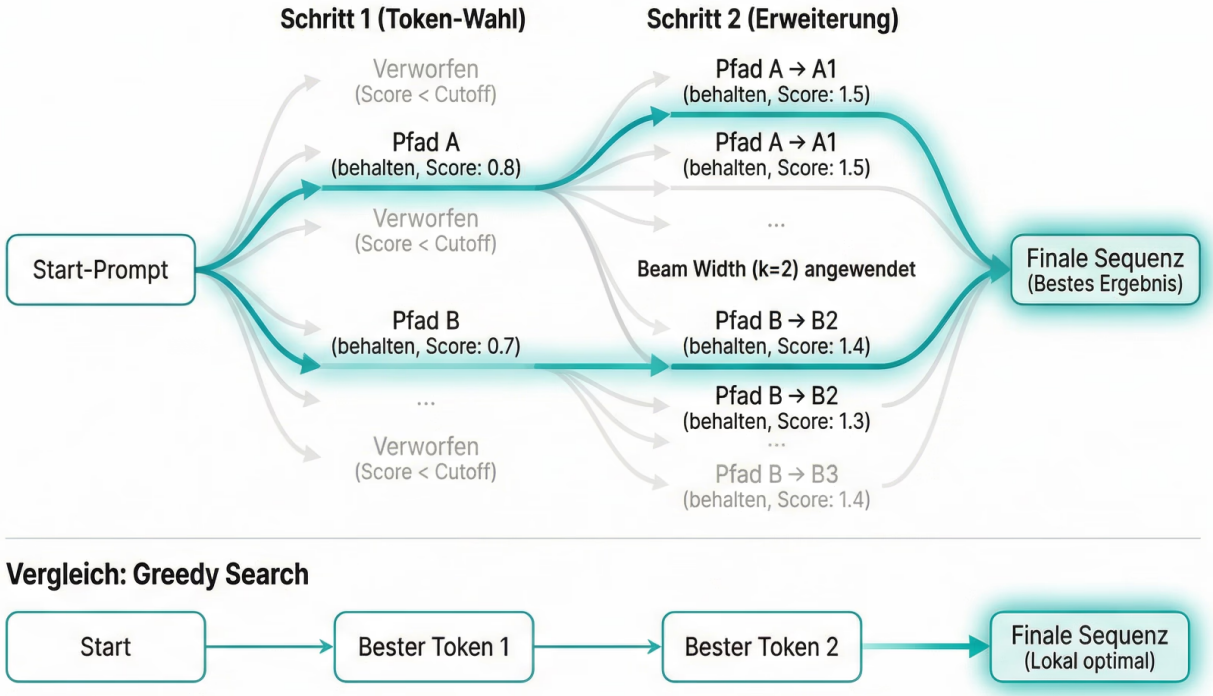
Was ist "Beam Search"?

Ein heuristischer Suchalgorithmus in der Sprachgenerierung (LLMs), der anstatt nur den besten nächsten Token (Greedy Search) auszuwählen, eine feste Anzahl ("Beam Width") der vielversprechendsten Sequenzen gleichzeitig verfolgt, um bessere, kohärentere Ergebnisse zu finden.

Breitere Suche ("Beam Width")
Verfolgt k beste Pfade parallel, nicht nur einen.

Bessere Ergebnisse
Findet oft kohärentere und global optimalere Sequenzen als Greedy Search.

Ressourcenabwägung
Höhere "Beam Width" verbessert Qualität, benötigt aber mehr Rechenleistung und Speicher.



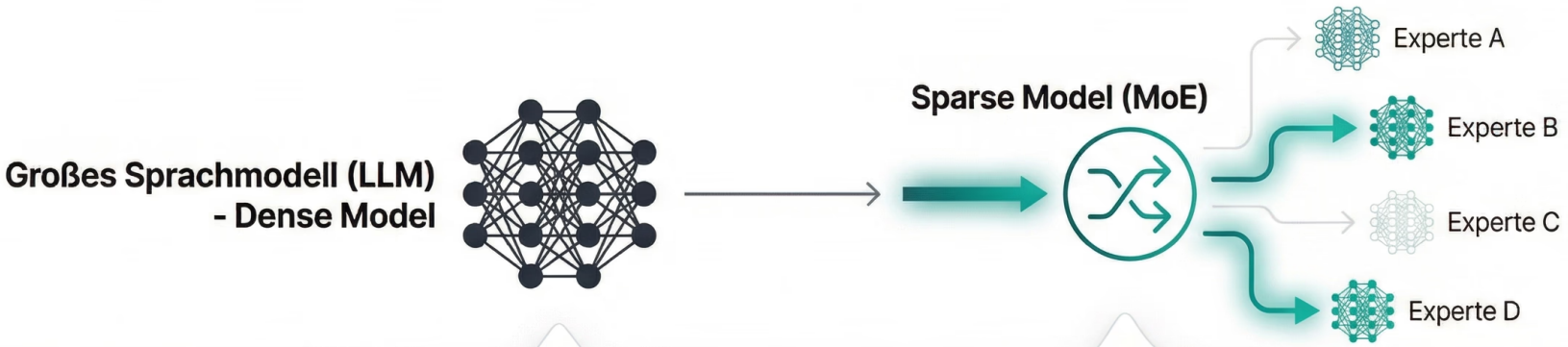
Kapitel 2.17

webconsulting

Kapitel 2.17: Was ist "Beam Search"?
webconsulting.at/blog/ki-kompendium-2025#217-was-ist-beam-search

Infografik: KI-generiert

Was sind "Sparse Models" (MoE)?



Kernkonzept: Mixture of Experts (MoE). Anstatt das gesamte Modell für jede Anfrage zu aktivieren, leitet ein "**Router**" die Eingabe nur an **relevante**, **spezialisierte Teilmodelle** (Experten) weiter. Dies reduziert die aktive Rechenlast erheblich.

Effizienz & Leistung

Schnellere Verarbeitung & geringere Kosten

Nur ein Bruchteil der Parameter wird pro Token aktiviert, was Training und Inferenz beschleunigt und Energie spart.

Skalierbarkeit

Ermöglicht riesige Modelle

MoE-Modelle können Billionen von Parametern umfassen, bleiben aber durch die selektive Aktivierung auf Standard-Hardware effizient ausführbar.

Spezialisierung

Fachwissen in Modulen

Verschiedene Experten lernen spezifische Aufgaben oder Domänen, wodurch das Modell insgesamt vielseitiger und präziser wird, ohne breites Wissen zu verlieren.

webconsulting

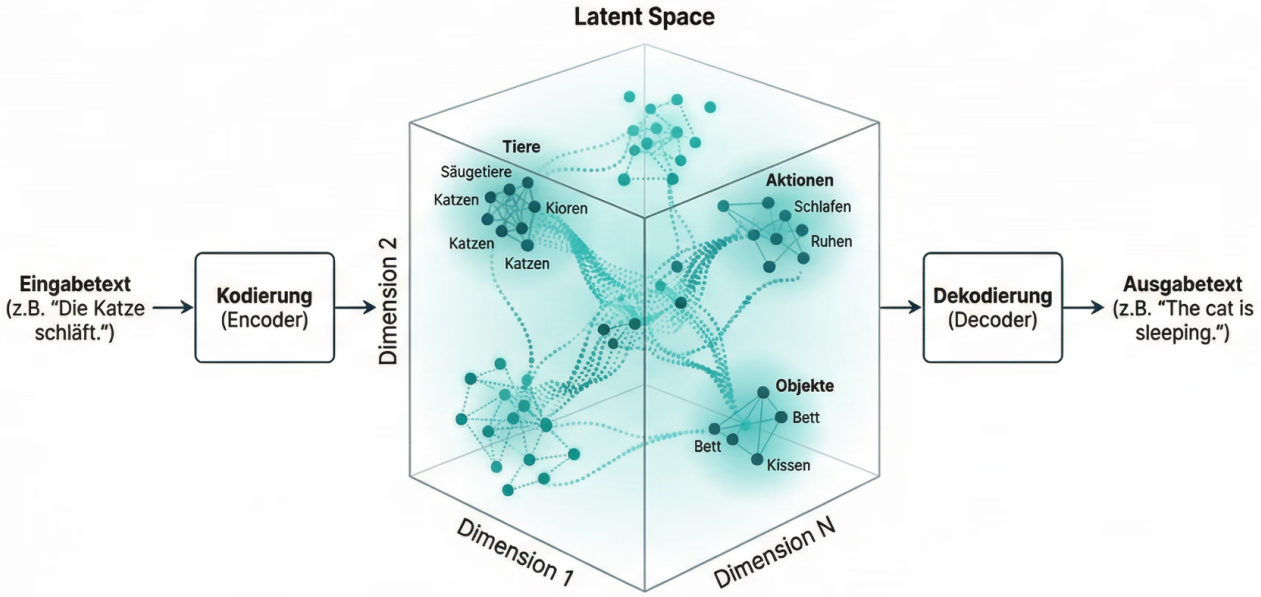
Kapitel 2.18: Was sind "Sparse Models" (MoE)?

webconsulting.at/blog/ki-kompendium-2025#218-was-sind-sparse-models-moe

Infografik: KI-generiert

Was ist "Latent Space"?

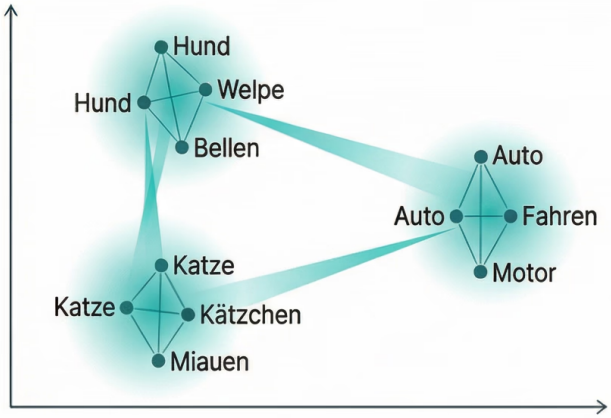
Der **"Latente Raum"** (Latent Space) ist ein mehrdimensionaler, konzeptioneller Raum, in dem ein Large Language Model (LLM) Informationen verarbeitet und speichert. Es ist eine interne Repräsentation von Wissen, in der ähnliche Konzepte und Bedeutungen räumlich nah beieinander liegen, kodiert als numerische Vektoren.



Vektorisierung von Bedeutung: Wörter, Sätze oder Konzepte werden in numerische Koordinaten (Vektoren) umgewandelt, die ihre Position im Raum bestimmen.

Semantische Nähe = Räumliche Nähe: Konzepte mit ähnlicher Bedeutung (z.B. "König" und "Königin") liegen im latenten Raum dicht beieinander.

Hochdimensionalität: Der Raum besteht aus Hunderten oder Tausenden von Dimensionen, um komplexe Zusammenhänge und Nuancen der Sprache zu erfassen.



Kapitel 2.19

webconsulting

Kapitel 2.19: Was ist "Latent Space"?

webconsulting.at/blog/ki-kompendium-2025#219-was-ist-latent-space

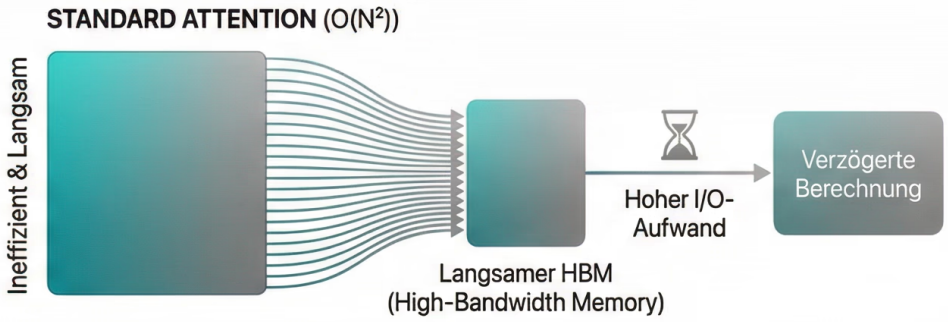
Infografik: KI-generiert

KAPITEL 2.20

Was ist “Flash Attention”?

Effizienz-Booster für Large Language Models (LLMs)

KERNELEMENT:
Eine optimierte Attention-Technologie, die Speichgriffe (I/O) drastisch reduziert, indem Berechnungen in kleinere Blöcke unterteilt und direkt im schnellen SRAM (GPU-Cache) ausgeführt werden, statt im langsamen HBM (High-Bandwidth Memory).



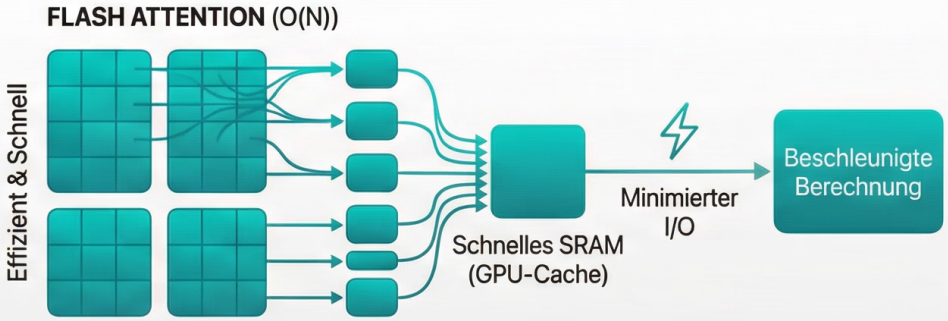
HAUPTVORTEILE

Schnellere Inferenz & Training
Beschleunigt Modellberechnungen signifikant durch minimierte Speicherlatenz.

Längere Kontextfenster
Ermöglicht die Verarbeitung von viel mehr Tokens, da der Speicherbedarf linear skaliert.

Geringerer Speicherverbrauch
Vermeidet die vollständige Materialisierung der großen Attention-Matrix, spart VRAM.

Skalierbarkeit für große Modelle
Essentiell für das Training moderner, massiver LLMs mit Milliarden von Parametern.



webconsulting

Kapitel 2.20: Was ist "Flash Attention"?
webconsulting.at/blog/ki-kompendium-2025#220-was-ist-flash-attention

Infografik: KI-generiert