DAV/DGVFM
Jahrestagung
2026

Zoran Nikolić & Sara Kristin Schmidt, Deloitte

Mathematische Grundlagen generativer KI

Fachgruppe ADS/AI, 30.04.2026

Fragen an MS Copilot (GPT-5)

Wie wird man Mitglied bei der DAV?

Um Mitglied beim **Deutschen Alpenverein (DAV)** zu werden, gehst du in der Regel in vier einfachen Schritten vor.[…]

Wie wird **eine Mathematikerin** Mitglied bei der DAV?

Um Mitglied der **Deutschen Aktuarvereinigung (DAV)** zu werden, folgt **eine Mathematikerin** in der Regel diesem Weg: […]

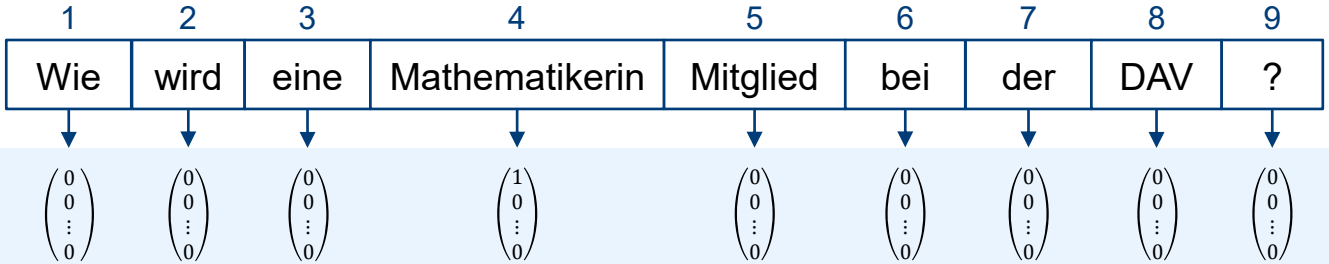
Wenn du mit DAV die **Deutsche Aktuarvereinigung e. V.** meinst (was im Kontext Mathematikerin am besten passt), dann läuft der Weg zur Mitgliedschaft über die Ausbildung zur Aktuarin DAV. […]

DAV Deutsche Aktuarvereinigung e.V.

DGVFM Deutsche Gesellschaft für Versicherungs- und Finanzmathematik e.V.

Tokenisierung

Vokabular ca. vom Umfang
$|\boldsymbol{V}| \approx 50\,Tsd.$

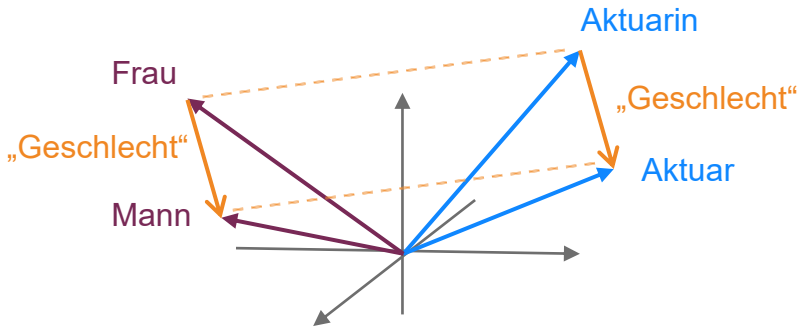


Repräsentation im $\mathbb{R}^{|\boldsymbol{V}|}$?

Embeddings

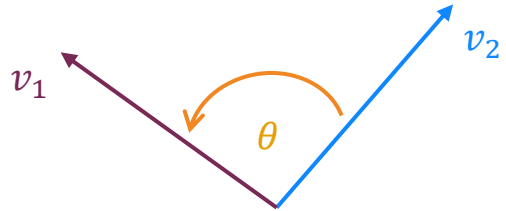
Repräsentation der Tokens als numerische Vektoren in „kleinerem“ Raum $\mathbb{R}^{4096}$ mit Hilfe von einer trainierbaren Embedding-Matrix $\boldsymbol{E} \in \mathbb{R}^{4096 \times 50.000}$, die das Folgende mathematisch erfasst:

Bedeutung von Tokens



$$Aktuar \approx Aktuarin + (Mann - Frau)$$

Ähnlichkeit von Tokens



$$Ähnlichkeit(v_1, v_2) := \frac{1}{|v_1||v_2|} v_1 \cdot v_2 = \cos(\theta)$$

DAV Deutsche Aktuarvereinigung e.V.

DGVFM Deutsche Gesellschaft für Versicherungs- und Finanzmathematik e.V.

Tokenisierung

Vokabular ca. vom Umfang
$|\boldsymbol{V}| = 50\,Tsd.$

1	2	3	4	5	6	7	8	9
Wie	wird	eine	Mathematikerin	Mitglied	bei	der	DAV	?

Embedding

Finden geeigneter
Repräsentationen im $\mathbb{R}^{4096}$

1	2	3	4	5	6	7	8	9
$\begin{pmatrix} 0{,}0107 \\ 0{,}0043 \\ \vdots \\ -0{,}0312 \end{pmatrix}$	$\begin{pmatrix} -0{,}0020 \\ 0{,}0237 \\ \vdots \\ -0{,}0194 \end{pmatrix}$	$\begin{pmatrix} 0{,}0483 \\ 0{,}0456 \\ \vdots \\ 0{,}0240 \end{pmatrix}$	$\begin{pmatrix} -0{,}0067 \\ -0{,}0348 \\ \vdots \\ -0{,}0082 \end{pmatrix}$	$\begin{pmatrix} 0{,}0236 \\ -0{,}0185 \\ \vdots \\ 0{,}0025 \end{pmatrix}$	$\begin{pmatrix} 0{,}0002 \\ 0{,}0159 \\ \vdots \\ -0{,}0109 \end{pmatrix}$	$\begin{pmatrix} 0{,}0489 \\ 0{,}0213 \\ \vdots \\ 0{,}0238 \end{pmatrix}$	$\begin{pmatrix} 0{,}0190 \\ 0{,}004 \\ \vdots \\ -0{,}0243 \end{pmatrix}$	$\begin{pmatrix} 0{,}0081 \\ -0{,}0033 \\ \vdots \\ 0{,}0194 \end{pmatrix}$

Transformerblöcke

Sukzessive Transformation der
Embeddings im $\mathbb{R}^{4096}$

DAV Deutsche Aktuarvereinigung e.V.

DGVFM Deutsche Gesellschaft für Versicherungs- und Finanzmathematik e.V.

Transformerblöcke

Idee: Schrittweise Verbesserung der Embeddings im $\mathbb{R}^{4096}$ durch sukzessives Durchlaufen der Transformerblöcke, wobei jeder Block alle Tokens parallel verarbeitet.

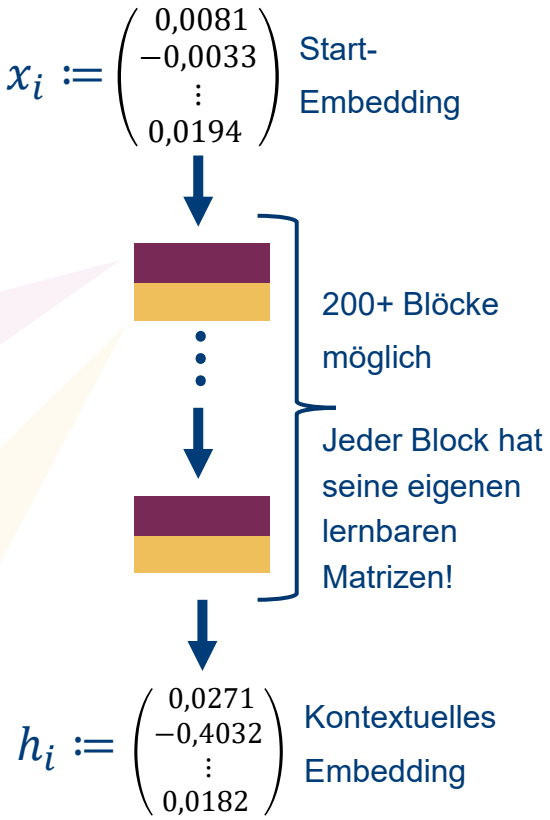
Multi-Head Attention: Jedes Token bezieht kontextuelle Informationen von anderen Tokens der Sequenz (gleich mehr!).

Feed-Forward Layer: Neuronales Netz für nicht-lineare Transformation, das von jedem Token i einzeln durchlaufen wird:

$$FFN(x_i) = W_2 \cdot \sigma(W_1 \cdot x_i)$$

Nicht-lineare Funktion

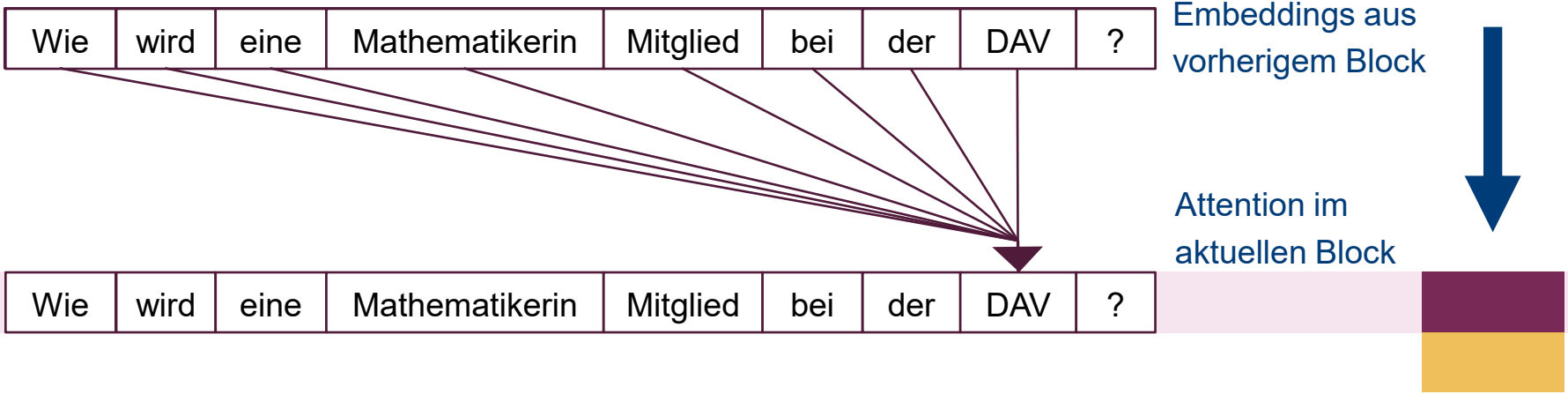
Lernbare Matrizen



Attention: Idee

Jedes Token bezieht kontextuelle Informationen von den anderen Tokens, die in der Sequenz davor liegen.

Beispielhaft für das Token **DAV** gilt:

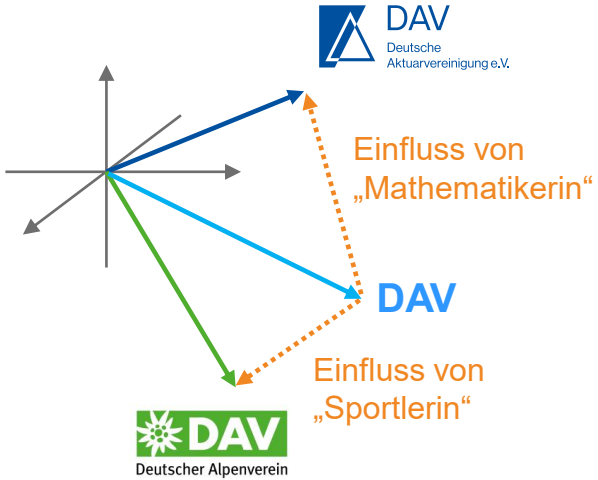


Multi-Head Attention: Mathematische Intuition

$$Head_i \approx \sum_{j=1}^{i} \underbrace{\left(\overbrace{(W^{Q} x_i)}^{\text{„Ich suche“}} \cdot \overbrace{(W^{K} x_j)}^{\text{„Ich biete“}} \right)}_{\text{Ähnlichkeit} \in \mathbb{R}} \cdot \underbrace{(W^{V} x_j)}_{\text{Anpassungs-richtung} \in \mathbb{R}^{4096}}$$

DAV Vektor

Mathematikerin Vektor



→ W^Q, W^K und W^V sind lernbare Matrizen des Attention Heads

Multi-Head Attention: Lineare Kombination mehrerer Heads (mit jeweils eigenen Matrizen), um der Komplexität der natürlichen Sprache gerecht zu werden.

DAV Deutsche Aktuarvereinigung e.V.

DGVFM Deutsche Gesellschaft für Versicherungs- und Finanzmathematik e.V.

Tokenisierung

Vokabular ca. vom Umfang
$|\boldsymbol{V}| = 50\,Tsd.$

1	2	3	4	5	6	7	8	9
Wie	wird	eine	Mathematikerin	Mitglied	bei	der	DAV	?

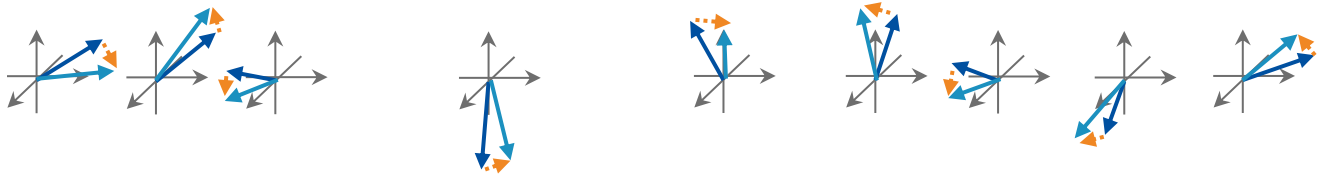
Embedding

Finden geeigneter
Repräsentationen im $\mathbb{R}^{4096}$

$\begin{pmatrix} 0{,}0107 \\ 0{,}0043 \\ \vdots \\ -0{,}0312 \end{pmatrix}$	$\begin{pmatrix} -0{,}0020 \\ 0{,}0237 \\ \vdots \\ -0{,}0194 \end{pmatrix}$	$\begin{pmatrix} 0{,}0483 \\ 0{,}0456 \\ \vdots \\ 0{,}0240 \end{pmatrix}$	$\begin{pmatrix} -0{,}0067 \\ -0{,}0348 \\ \vdots \\ -0{,}0082 \end{pmatrix}$	$\begin{pmatrix} 0{,}0236 \\ -0{,}0185 \\ \vdots \\ 0{,}0025 \end{pmatrix}$	$\begin{pmatrix} 0{,}0002 \\ 0{,}0159 \\ \vdots \\ -0{,}0109 \end{pmatrix}$	$\begin{pmatrix} 0{,}0489 \\ 0{,}0213 \\ \vdots \\ 0{,}0238 \end{pmatrix}$	$\begin{pmatrix} 0{,}0190 \\ 0{,}004 \\ \vdots \\ -0{,}0243 \end{pmatrix}$	$\begin{pmatrix} 0{,}0081 \\ -0{,}0033 \\ \vdots \\ 0{,}0194 \end{pmatrix}$

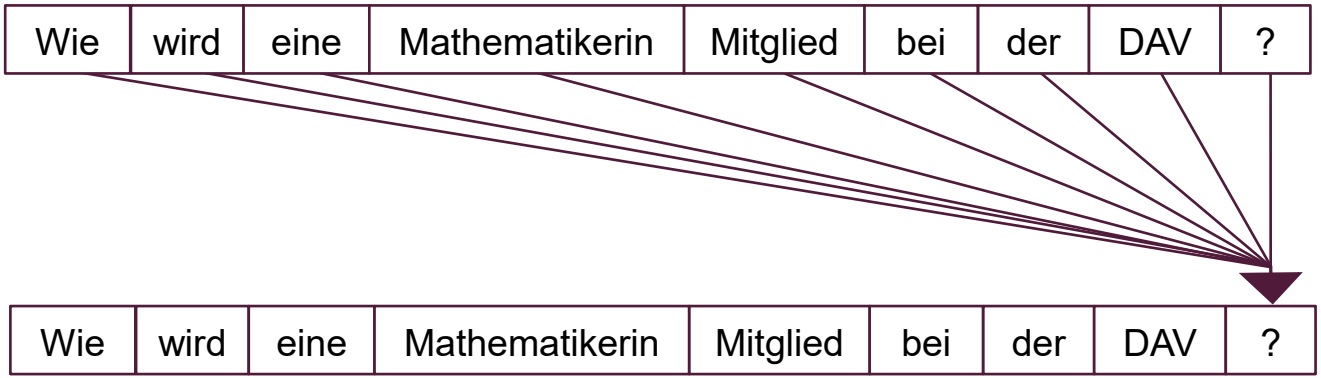
Transformerblöcke

Sukzessive Transformation der
Embeddings im $\mathbb{R}^{4096}$



Transformerblöcke: Output

Das letzte Token bezieht kontextuelle Informationen von **allen** anderen Tokens. Es genügt daher, aus dem Output des letzten Transformerblocks nur das kontextuelle Embedding des letzten Tokens weiterzuverarbeiten.



DAV Deutsche Aktuarvereinigung e.V.

DGVFM Deutsche Gesellschaft für Versicherungs- und Finanzmathematik e.V.

Tokenisierung

Vokabular ca. vom Umfang
$|\boldsymbol{V}| = 50\,Tsd.$

1	2	3	4	5	6	7	8	9
Wie	wird	eine	Mathematikerin	Mitglied	bei	der	DAV	?

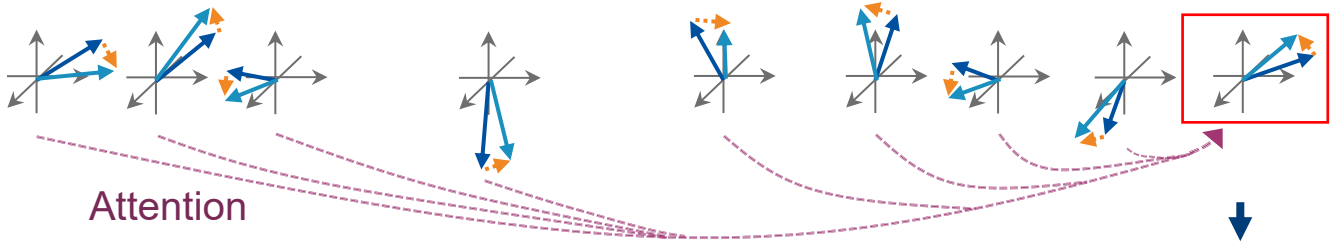
Embedding

Finden geeigneter
Repräsentationen im $\mathbb{R}^{4096}$

$$\begin{pmatrix} 0{,}0107 \\ 0{,}0043 \\ \vdots \\ -0{,}0312 \end{pmatrix} \quad \begin{pmatrix} -0{,}0020 \\ 0{,}0237 \\ \vdots \\ -0{,}0194 \end{pmatrix} \quad \begin{pmatrix} 0{,}0483 \\ 0{,}0456 \\ \vdots \\ 0{,}0240 \end{pmatrix} \quad \begin{pmatrix} -0{,}0067 \\ -0{,}0348 \\ \vdots \\ -0{,}0082 \end{pmatrix} \quad \begin{pmatrix} 0{,}0236 \\ -0{,}0185 \\ \vdots \\ 0{,}0025 \end{pmatrix} \quad \begin{pmatrix} 0{,}0002 \\ 0{,}0159 \\ \vdots \\ -0{,}0109 \end{pmatrix} \quad \begin{pmatrix} 0{,}0489 \\ 0{,}0213 \\ \vdots \\ 0{,}0238 \end{pmatrix} \quad \begin{pmatrix} 0{,}0190 \\ 0{,}004 \\ \vdots \\ -0{,}0243 \end{pmatrix} \quad \begin{pmatrix} 0{,}0081 \\ -0{,}0033 \\ \vdots \\ 0{,}0194 \end{pmatrix}$$

Transformerblöcke

Sukzessive Transformation der
Embeddings im $\mathbb{R}^{4096}$



DAV Deutsche Aktuarvereinigung e.V.

DGVFM Deutsche Gesellschaft für Versicherungs- und Finanzmathematik e.V.

Language Modeling Head

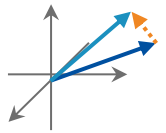
Unembedding

Projektion des „komprimierten“ Vektors im $\mathbb{R}^{4096}$ zurück auf den größeren Raum $\mathbb{R}^{|V|} = \mathbb{R}^{50.000}$ mittels trainierbarer Unembedding-Matrix $U \in \mathbb{R}^{50.000 \times 4096}$.

Softmax-Funktion

Ermittlung der Wahrscheinlichkeitsverteilung über alle Tokens für das nächste zu wählende Token bedingt auf bisherige Token-Sequenz als

$$p_i = \frac{\exp(u_i)}{\sum_{j=1}^{|V|} \exp(u_j)} \qquad \forall i \in 1, \dots, |V|.$$



$$h_9 := \begin{pmatrix} 0{,}0271 \\ -0{,}4032 \\ \vdots \\ 0{,}0182 \end{pmatrix} \in \mathbb{R}^{4096}$$

$$u := \begin{pmatrix} -0{,}3578 \\ 4{,}1241 \\ \vdots \\ -2{,}0765 \end{pmatrix} \in \mathbb{R}^{|V|}$$

$$p := \begin{pmatrix} 0{,}0001 \\ 0{,}1240 \\ \vdots \\ 0{,}0000 \end{pmatrix} \begin{matrix} \textbf{Mathematiker} \\ \textbf{Um} \\ \\ \textbf{Sonnig} \end{matrix}$$

DAV Deutsche Aktuarvereinigung e.V.

DGVFM Deutsche Gesellschaft für Versicherungs- und Finanzmathematik e.V.

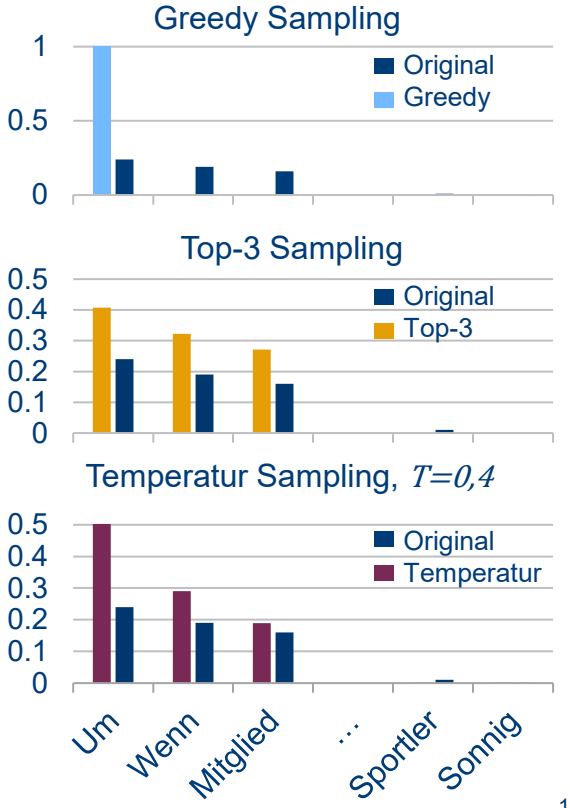
Sampling

Greedy Sampling: Wähle das Token mit der höchsten Wahrscheinlichkeit.

Top-k Sampling ($k \in \mathbb{N}$): Ziehe aus der Menge der k Tokens mit den k höchsten Wahrscheinlichkeiten gemäß ihrer (reskalierten) Wahrscheinlichkeiten.

Temperatur Sampling: Modifiziere die Softmax-Funktion mittels des Temperatur-Parameters $T \in \mathbb{R}^+$ als

$$p_i = \frac{\exp\left(\frac{u_i}{T}\right)}{\sum_{j=1}^{|V|} \exp\left(\frac{u_j}{T}\right)} \qquad \forall i \in 1, \dots, |V|.$$



DAV Deutsche Aktuarvereinigung e.V.

DGVFM Deutsche Gesellschaft für Versicherungs- und Finanzmathematik e.V.

Tokenisierung

Vokabular ca. vom Umfang $|\boldsymbol{V}| = 50\,Tsd.$

1	2	3	4	5	6	7	8	9
Wie	wird	eine	Mathematikerin	Mitglied	bei	der	DAV	?

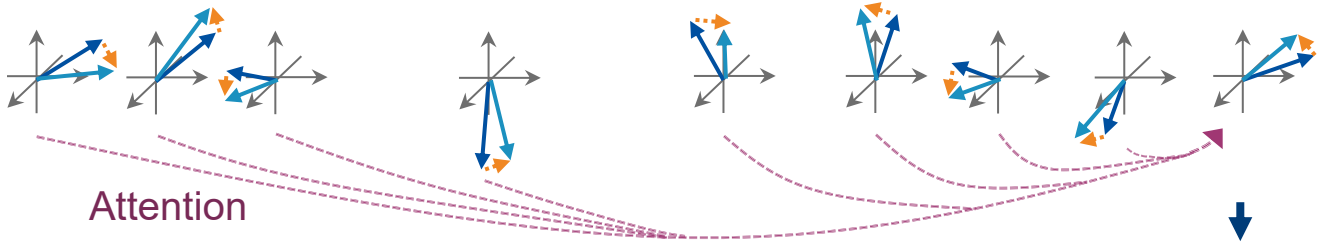
Embedding

Finden geeigneter Repräsentationen im $\mathbb{R}^{4096}$

$\begin{pmatrix} 0{,}0107 \\ 0{,}0043 \\ \vdots \\ -0{,}0312 \end{pmatrix}$	$\begin{pmatrix} -0{,}0020 \\ 0{,}0237 \\ \vdots \\ -0{,}0194 \end{pmatrix}$	$\begin{pmatrix} 0{,}0483 \\ 0{,}0456 \\ \vdots \\ 0{,}0240 \end{pmatrix}$	$\begin{pmatrix} -0{,}0067 \\ -0{,}0348 \\ \vdots \\ -0{,}0082 \end{pmatrix}$	$\begin{pmatrix} 0{,}0236 \\ -0{,}0185 \\ \vdots \\ 0{,}0025 \end{pmatrix}$	$\begin{pmatrix} 0{,}0002 \\ 0{,}0159 \\ \vdots \\ -0{,}0109 \end{pmatrix}$	$\begin{pmatrix} 0{,}0489 \\ 0{,}0213 \\ \vdots \\ 0{,}0238 \end{pmatrix}$	$\begin{pmatrix} 0{,}0190 \\ 0{,}004 \\ \vdots \\ -0{,}0243 \end{pmatrix}$	$\begin{pmatrix} 0{,}0081 \\ -0{,}0033 \\ \vdots \\ 0{,}0194 \end{pmatrix}$

Transformerblöcke

Sukzessive Transformation der Embeddings im $\mathbb{R}^{4096}$



Language Modeling Head

Projektion in $\mathbb{R}^{|V|}$, Umwandlung in Wahrscheinlichkeiten

$$\begin{pmatrix} 0{,}0001 \\ 0{,}1240 \\ \vdots \\ 0{,}0000 \end{pmatrix}$$

Sampling

Um

Take Home Messages

Architektur der Sprachmodelle:

- Basiert auf überraschend einfachen mathematischen Methoden der Linearen Algebra (Vektoren, Matrixmultiplikation), Stochastik (Sampling) und Analysis (Gradientenabstieg im Training).
- Durchbruch im letzten Jahrzehnt durch den Attention-Mechanismus: „Mathematikerin“ im Beispielsatz beeinflusst die Bedeutung von „DAV“.
- Neueste Modelle besitzen mehr als 1 Billion Parameter: Extrem eingeschränkte Interpretierbarkeit.

Hinweise zur Anwendung im Aktuariat:

- **Don´t:** Anwendungen, bei denen exakte, reproduzierbare und nachvollziehbare Antworten nötig sind.
- **Do:** Text-Zusammenfassungen, Modell-Dokumentationen, Code-Erklärungen und Code-Erstellung.

DAV Deutsche Aktuarvereinigung e.V.

DGVFM Deutsche Gesellschaft für Versicherungs- und Finanzmathematik e.V.

Zeit für Fragen!

Für Interessierte:

Fit4AI kompakt

Nächste Termine:

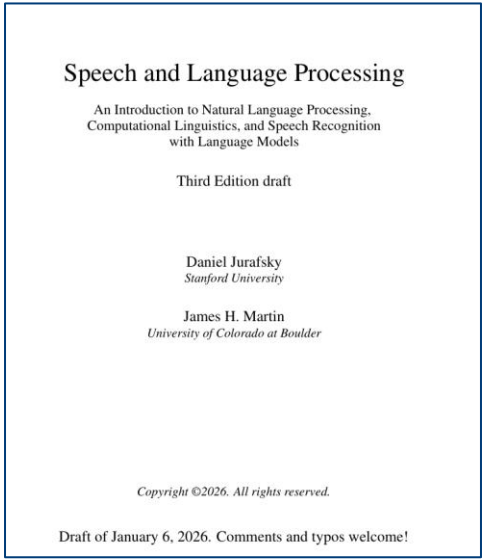
21.-22.09.2026

13.-14.10.2026

03.-04.11.2026

07.-08.12.2026

Veranstaltungen für Aktuare - Deutsche Aktuarvereinigung e.V.



Speech and Language Processing

An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models

Third Edition draft

Daniel Jurafsky
Stanford University

James H. Martin
University of Colorado at Boulder

Copyright ©2026. All rights reserved.

Draft of January 6, 2026. Comments and typos welcome!

https://web.stanford.edu/~jurafsky/slp3

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

https://arxiv.org/pdf/1706.03762

DAV/DGVFM
Jahrestagung
2026

Vielen Dank für Ihre Aufmerksamkeit.