Das disruptive Potential von KI in unserer Zeit

Osnabrücker Wissenschaftliche Gesellschaft (OWiG)
Osnabrück – 23.04.2025

Kai-Uwe Kühnberger

Überblick

- Was ist alles KI und wann begann KI?
- Beispiele des Potentials von Transformer-Architekturen
- Wie unterscheiden sich Transformer von anderen KI-Ansätzen?
  - Architektur, „Attention“, „Position System“ / Skalierung
- Disruptive Trends in der KI-Entwicklung
  - Knowledge Distillation oder wie trainiere ich ein kleines Netz?
  - Transfer von Wissen
- Möglichkeiten der Anwendung von KI-Technologie

Wo ist alles KI und wann begann KI?

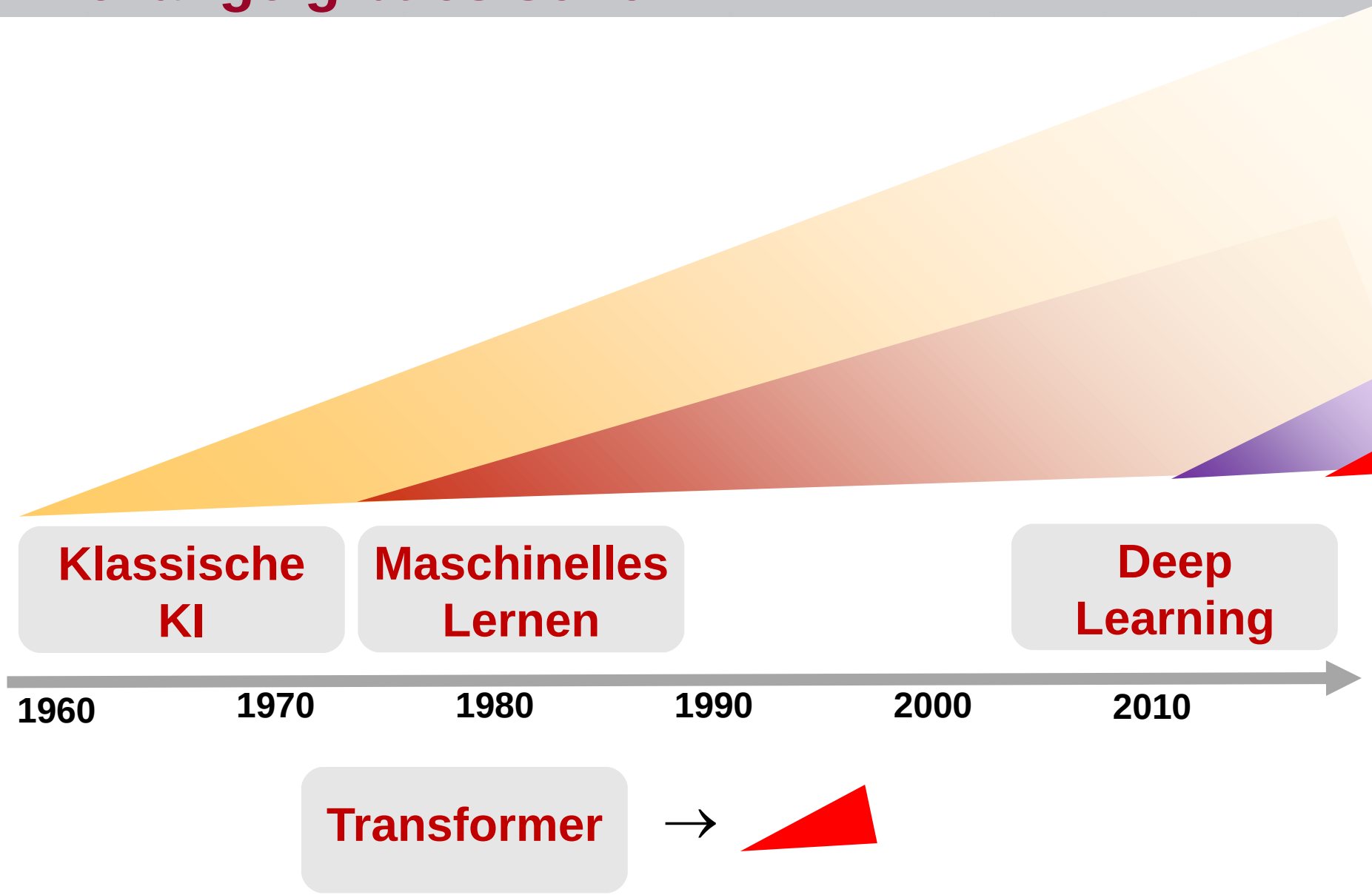
- Was ist KI?
- Eine minimale Geschichte der KI
- Was ist an der gegenwärtigen Situation neu?

Was ist Künstliche Intelligenz?

- KI ist die Simulation menschlicher Intelligenz durch Computerprogramme oder Roboter.

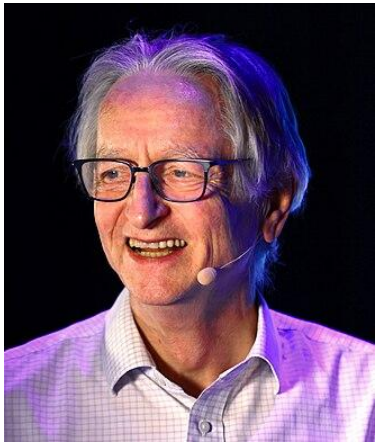
- KI simuliert Fähigkeiten wie
  - Entscheidungen fällen
  - Strategien verfolgen
  - Wahrnehmen
  - motorische Handlungen durchführen
  - abstrakte Konzepte lernen
  - Sprache benutzen
  - Musik komponieren oder spielen
  - Etc.

Wie lange gibt es schon KI?



Einige Meilensteine der KI

Nobel-Preise 2024 für Physik und Chemie



Geoffrey Hinton

John Hopfield

- Nobel-Preise für Physik 2024
  - Geoffrey Hinton
  - John Hopfield
- Nobel-Preise für Chemie 2024:
  - John Jumper
  - Demis Hassabis
  - [David Baker]

John Jumper

Demis Hassabis

- Hinton hat keinen Bezug zur Physik, Hassabis keinen zur Chemie, alle vier Genannten sind Wissenschaftler der KI
- Hinton, Hassabis und Jumper arbeiten bzw. arbeiteten für Google

Hype um KI

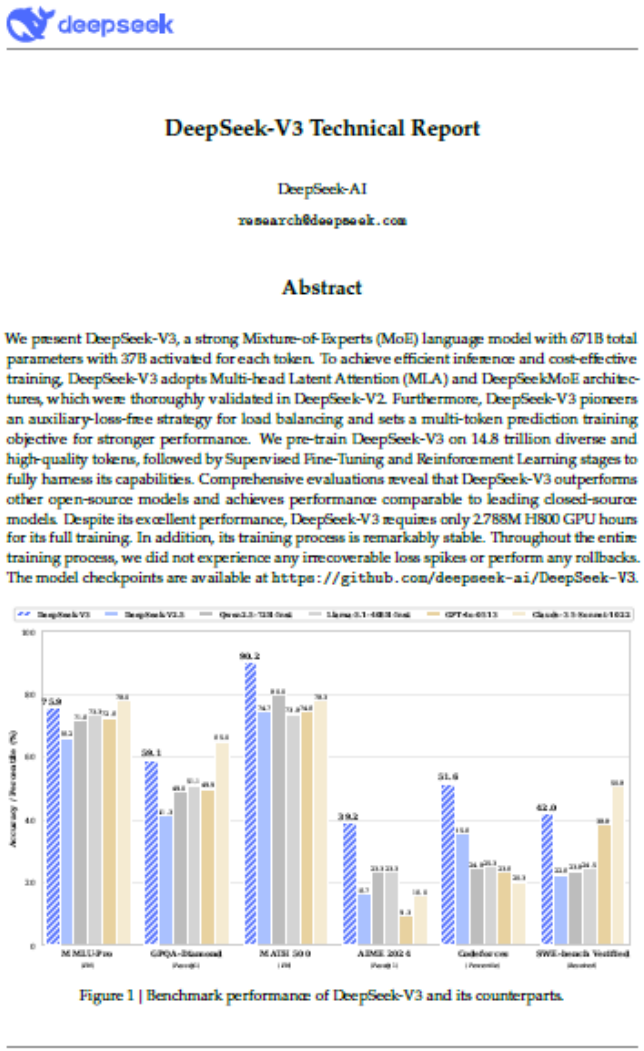
- Die Entscheidung Hinton, Hopfield, Hassabis und Jumper den Nobelpreis zuzuerkennen, ist nicht unumstritten.
- Interessant ist, dass die Nobelpreisträger mit Transformer-Architekturen und generativer KI wie ChatGPT relativ wenig zu tun haben.
- Was ist neu an GPT-Modellen, an Gemini, Llama etc.?
  - Kontext / Gedächtnis, Multimodalität, Transfer Learning, Skalierung, General Purpose Models etc.
- Transformer-Modelle (z.B. GPT-Modelle) sind eine KI-Technologie, die der AGI Idee („Artificial General Intelligence“) signifikant näher kommen.

Beispiele des Potentials von Transformer-Architekturen

- Beispiele für Anwendungen von Transformern

Beispiel: Automatisch erzeugte Podcasts

- Google bietet einen Service an, der einen mehr oder weniger beliebigen Text automatisiert in einen Podcast verwandelt.
  - Input: Technical Paper DeepSeek-V3.
  - Output: Podcast, der die Inhalte zusammenfasst.



deepseek

DeepSeek-V3 Technical Report

DeepSeek-AI
research@deepseek.com

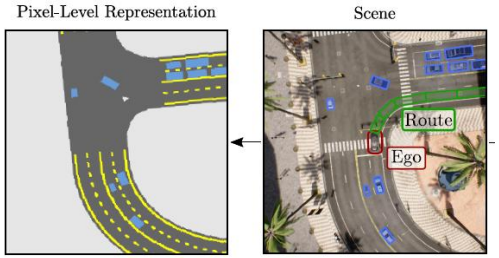
Abstract

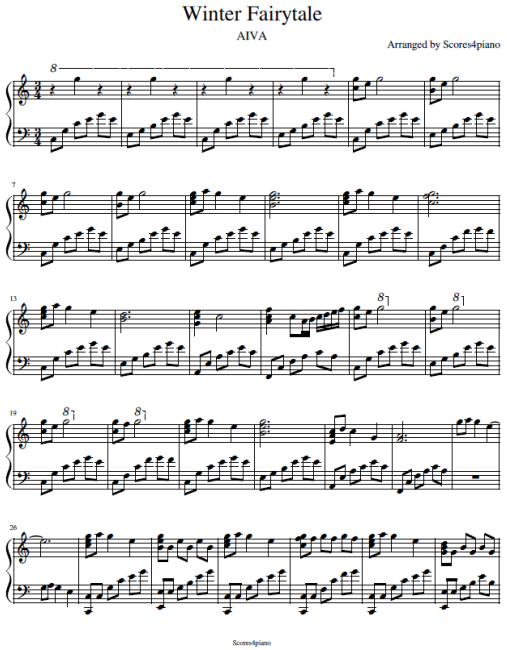
We present DeepSeek-V3, a strong Mixture-of-Experts (MoE) language model with 671B total parameters with 37B activated for each token. To achieve efficient inference and cost-effective training, DeepSeek-V3 adopts Multi-head Latent Attention (MLA) and DeepSeekMoE architectures, which were thoroughly validated in DeepSeek-V2. Furthermore, DeepSeek-V3 pioneers an auxiliary-loss-free strategy for load balancing and sets a multi-token prediction training objective for stronger performance. We pre-train DeepSeek-V3 on 14.8 trillion diverse and high-quality tokens, followed by Supervised Fine-Tuning and Reinforcement Learning stages to fully harness its capabilities. Comprehensive evaluations reveal that DeepSeek-V3 outperforms other open-source models and achieves performance comparable to leading closed-source models. Despite its excellent performance, DeepSeek-V3 requires only 2.788M H800 GPU hours for its full training. In addition, its training process is remarkably stable. Throughout the entire training process, we did not experience any irrecoverable loss spikes or perform any rollbacks. The model checkpoints are available at https://github.com/deepseek-ai/DeepSeek-V3.

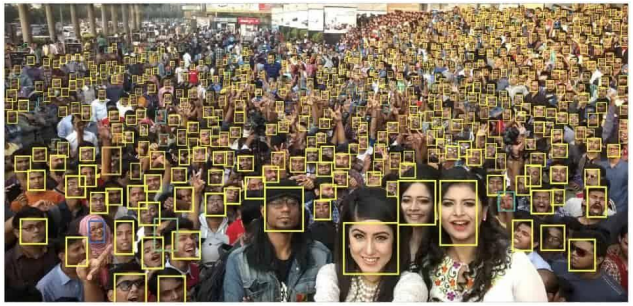
Figure 1 | Benchmark performance of DeepSeek-V3 and its counterparts.

Weitere Beispiele

- Transformer können noch mehr:
  - Bilder malen / Objekte erkennen
  - Gedichte schreiben
  - Inhalte zusammenfassen
  - Programmieren
  - Pläne erzeugen
  - Texte übersetzen / Texte überarbeiten
  - Etc.







```
Prolog                                                    Kopieren

magic_square([A1, A2, A3, A4,
              B1, B2, B3, B4,
              C1, C2, C3, C4,
              D1, D2, D3, D4]) :-
    permutation([1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16],
                [A1, A2, A3, A4, B1, B2, B3, B4, C1, C2, C3, C4, D1, D2, D3,
    A1 + A2 + A3 + A4 =:= 34,
    B1 + B2 + B3 + B4 =:= 34,
    C1 + C2 + C3 + C4 =:= 34,
    D1 + D2 + D3 + D4 =:= 34,
    A1 + B1 + C1 + D1 =:= 34,
    A2 + B2 + C2 + D2 =:= 34,
    A3 + B3 + C3 + D3 =:= 34,
    A4 + B4 + C4 + D4 =:= 34,
    A1 + B2 + C3 + D4 =:= 34,
    A4 + B3 + C2 + D1 =:= 34.
```

Beispiel: Multimodale Verarbeitung

- Multimodale Prozesse:
  - Input: Bild, Output: Gedicht

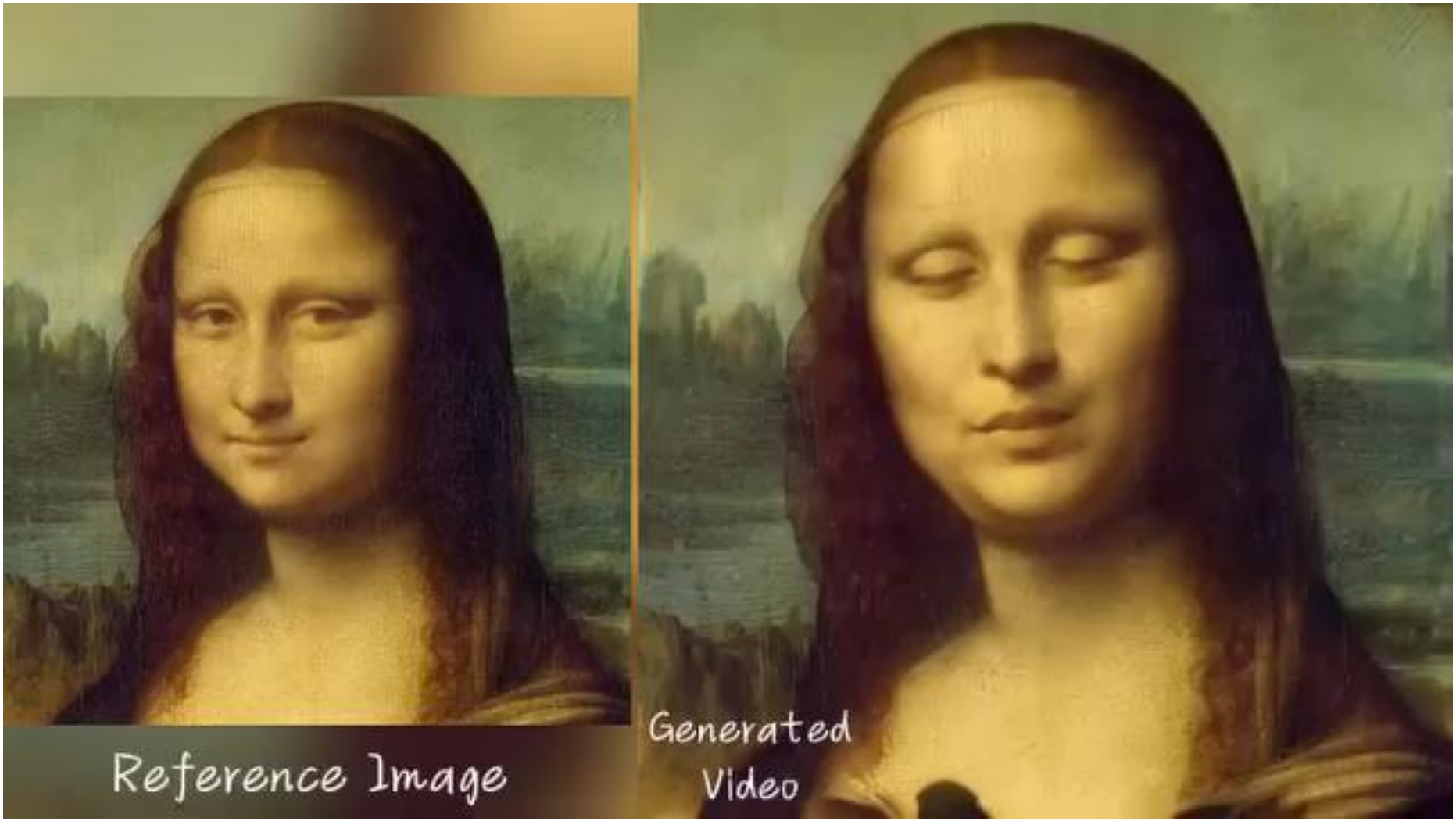
Drei Freundinnen lehnen aus dem Fenster,
Mit Gläsern in der Hand, so heiter.
Der Tag ist hell, die Sonne lacht,
Ein Moment, der Freude macht.

Sie stoßen an auf gute Zeiten,
Auf Freundschaft, die sie stets begleiten.
Ein Lächeln hier, ein Lachen dort,
In diesem Augenblick zählt kein Ort.

Die Welt da draußen mag sich drehen,
Doch hier im Fenster bleibt sie stehen.
Ein Schluck, ein Blick, ein stilles Wort,
Freundschaft trägt uns immer fort.

Beispiel: Video

- Erzeugung von Videos aus einem einzigen Bild



Wie unterscheiden sich Transformer von anderen KI-Ansätzen?

- Architektur
- Positional Encoding / Skalierung
- Attention

Was ist neu an Transformern?

- Erstens: es handelt sich um eine smarte und (sehr) komplexe Architektur:
  - **Transformer-Architektur**
  - **Es gibt verschiedene Transformer-Architekturen**

- Zweitens: durch Kodierung der Position eines Wortes im Satz („Positional Encoding“) wird das Training parallelisiert.
  - Dadurch können um Größenordnungen mehr Daten prozessiert werden ohne die Positions eines Tokens in einem Input zu verlieren.

- Drittens: ein „Attention“-Mechanismus kann das Problem des Kontexts (des Gedächtnis) in gewisser Weise lösen.

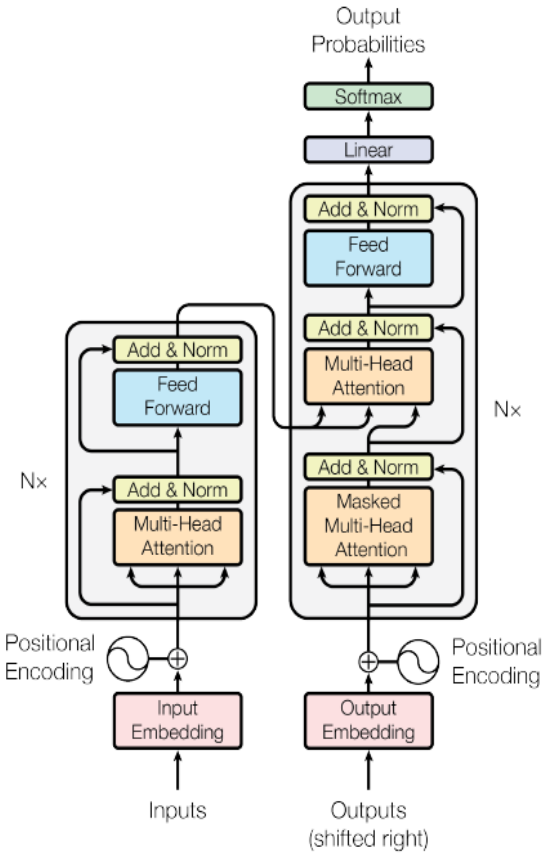


Figure 1: The Transformer - model architecture.

Vaswani et al. (2017)

Intuitive Idee von (Self-)Attention

- (Self-)Attention versucht relevante Elemente im eigenen Input zu identifizieren: Dadurch wird ein Kontext erzeugt.
  - Interessant ist, dass zunächst die lokale Umgebung analysiert wird und erst später die globalere Umgebung.

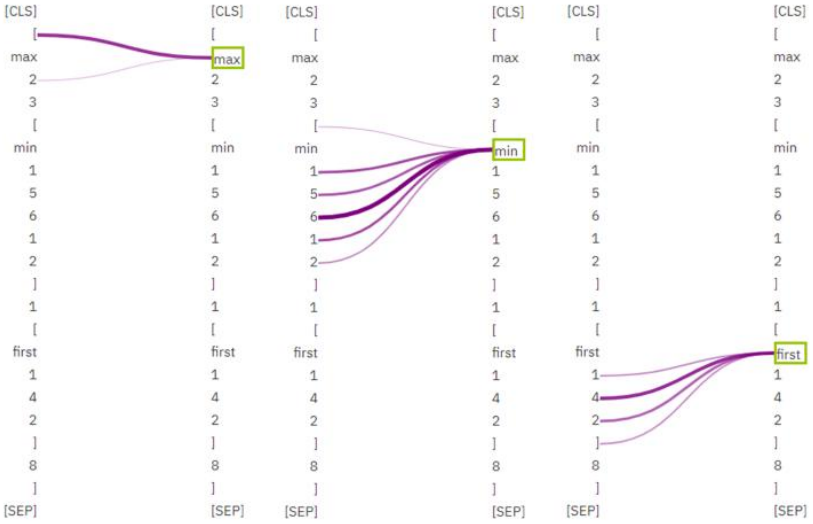


FIGURE 3.4: Figure shows that operators inside each sub-sequence are attended by their tokens in layer 6.

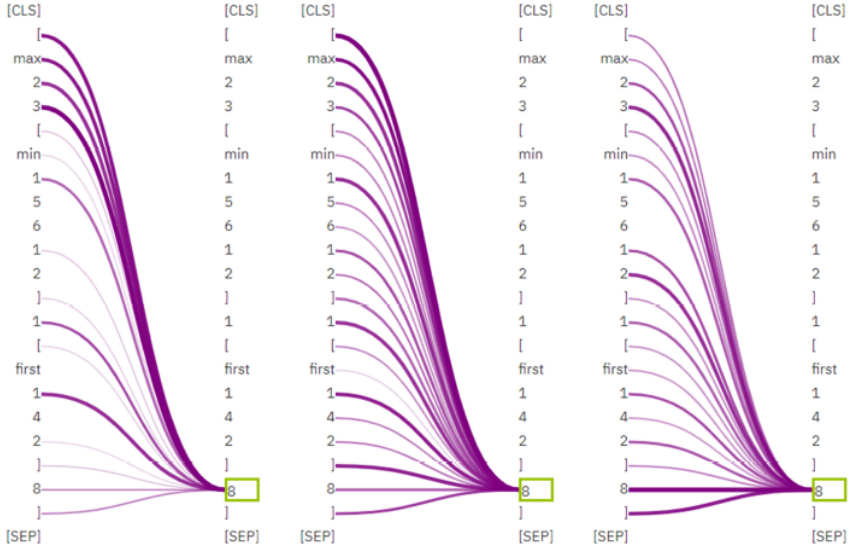


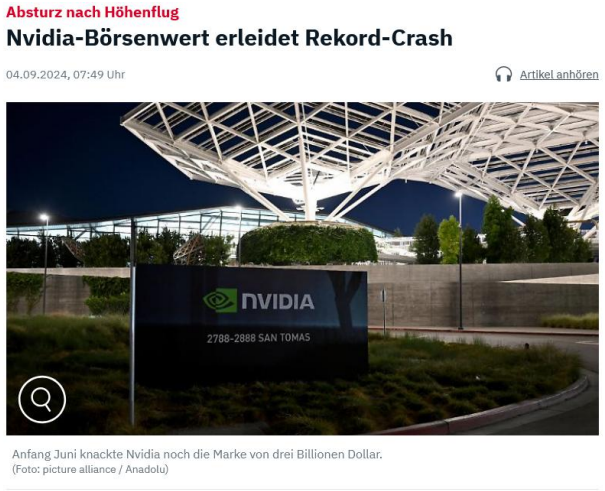
FIGURE 3.5: Figure shows that most of the tokens attend to the correct answer in layers 10, 11, and 12 from left to right

Ballout et al. (2023a)

Vorteile der Transformer Architekturen

- Warum Transformer-Architekturen disruptiv sind:
  - Transformer modellieren Kontext und Gedächtnis
  - Transformer können vortrainiert werden
  - Transformer realisieren Transfer-Learning
  - Transformer können skaliert werden
  - Transformer können mit multi-modalen Daten trainiert werden
  - Zukünftige Entwicklung: Transformer können Reasoning (Chain-of-Thought), Mathematik und Abstraktion realisieren (z.B. GPT-4o, DeepSeek-V3, DeepSeek-R1).
  - Zukünftige Entwicklung: Transformer können auch mit hierarchischem Wissen trainiert werden -> Hybride Architektur

Der 2 Billionen Dollar Crash



- Chinesisches Start-Up „DeepSeek“ hat kostengünstiges Sprachmodell entwickelt, das performant ist, aber weniger Energie, weniger Parameter, weniger Training, weniger Ressourcen benötigt.

- Was macht DeepSeek anders im Vergleich zu anderen Sprachmodellen?

Performanz von DeepSeek-V3

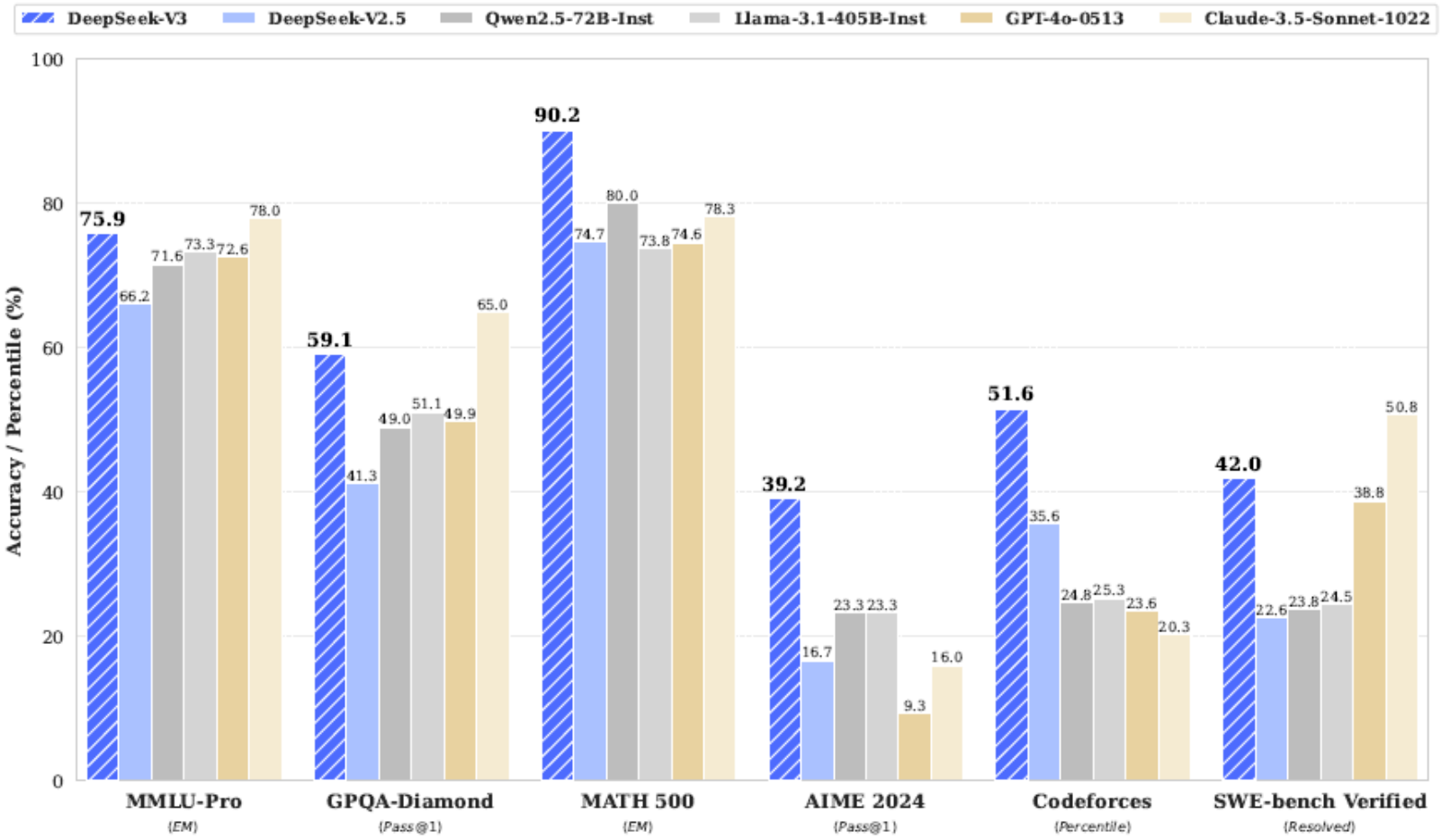


Figure 1 | Benchmark performance of DeepSeek-V3 and its counterparts.

DeepSeek-V3 Technical Report (27.12.2024)

Performanz von DeepSeek-R1

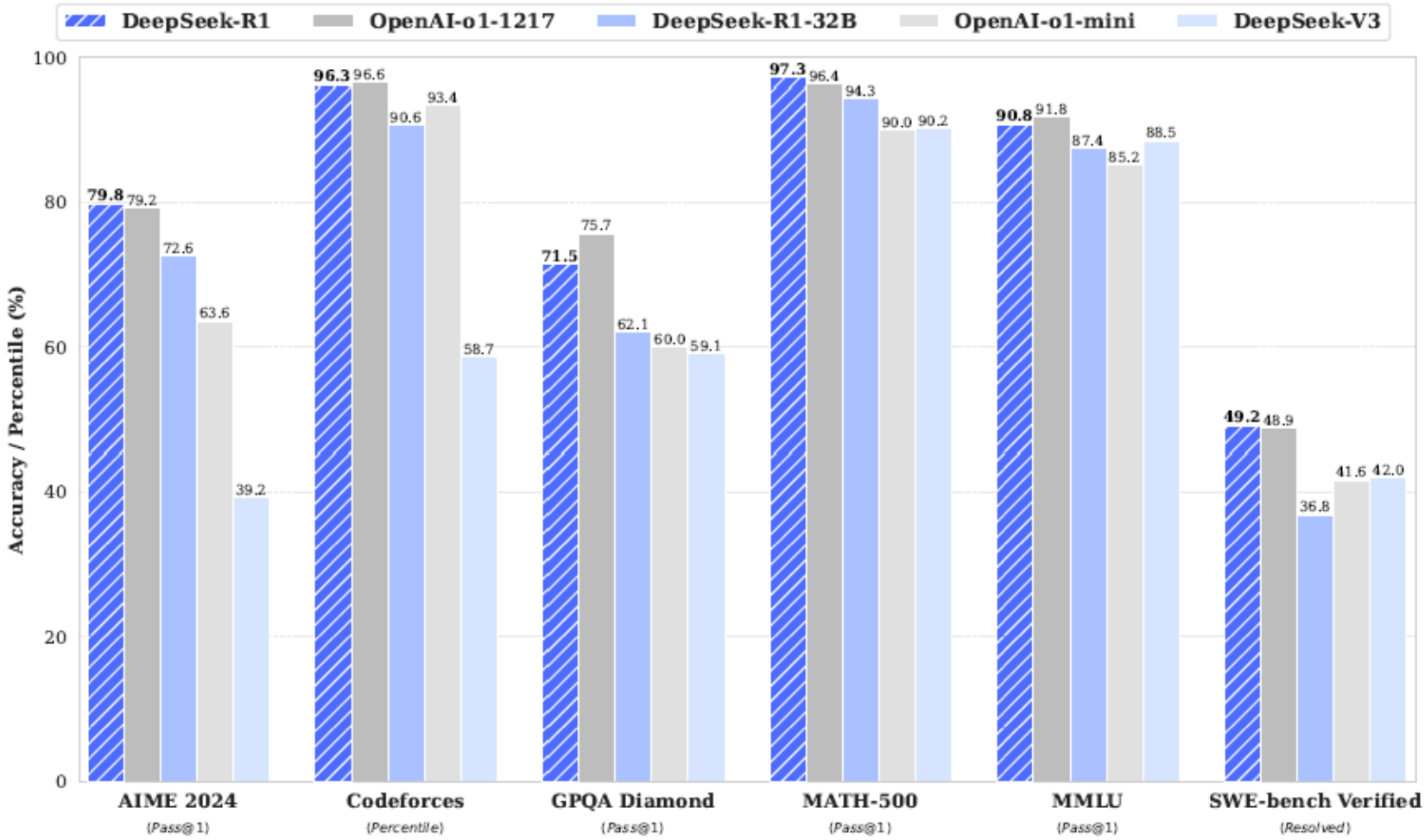


Figure 1 | Benchmark performance of DeepSeek-R1.

DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via RL (22.01.2025)

Was macht DeepSeek-V3 anders?

- Architektur:
  - Optimiertes „Positional Embedding“ (RoPE) und komprimierte Attention Queries
  - Trade-off zwischen Auslastung der Sub-Modelle und Performanz ist verbessert („Mixture of Experts Modell“)
  - Verbesserte Multi-Token Prediction
  - Etc.
- Post-Traning
  - „Knowledge Distillation“ DeepSeek-R1_Distill_DeepSeek-V3

- **Kosten des Trainings von ca. 5,6 Mio. Euro**

Was macht DeepSeek-R1 anders?

- Post-Training
  - Large-Scale Reinforcement Learning: Chain-of-Thought wird exploriert
  - Supervised Fine-Tuning
- Knowledge Distillation
  - Trainierte kleine Modelle (DeepSeek-R1_Distill_Qwen-32B) sind bezüglich der Performanz im Bereich von OpenAI-o1-mini.
- Kosten von R1: unklar, vermutlich hoch

- **Kleine Modelle können durch „Knowledge Distillation“ durch große Modelle effizient und ohne viele Ressourcen trainiert werden (mit einer vergleichbaren Performanz)**
- **Soweit ich das sehe sind keine neuen Methoden vorgeschlagen worden.**

Disruptive Trends in der KI-Entwicklung

Knowledge Distillation…

… oder wie trainiere ich ein kleines Netz?

Gegenwärtige Situation

- Thesen zur gegenwärtigen Situation
  - Die großen Modelle werden jetzt und in Zukunft außerhalb Europas entwickelt (Rechenkapazität, Energie, Kosten etc.)

  - Große und leistungsstarke Modelle werden auch in Zukunft proprietär sein.

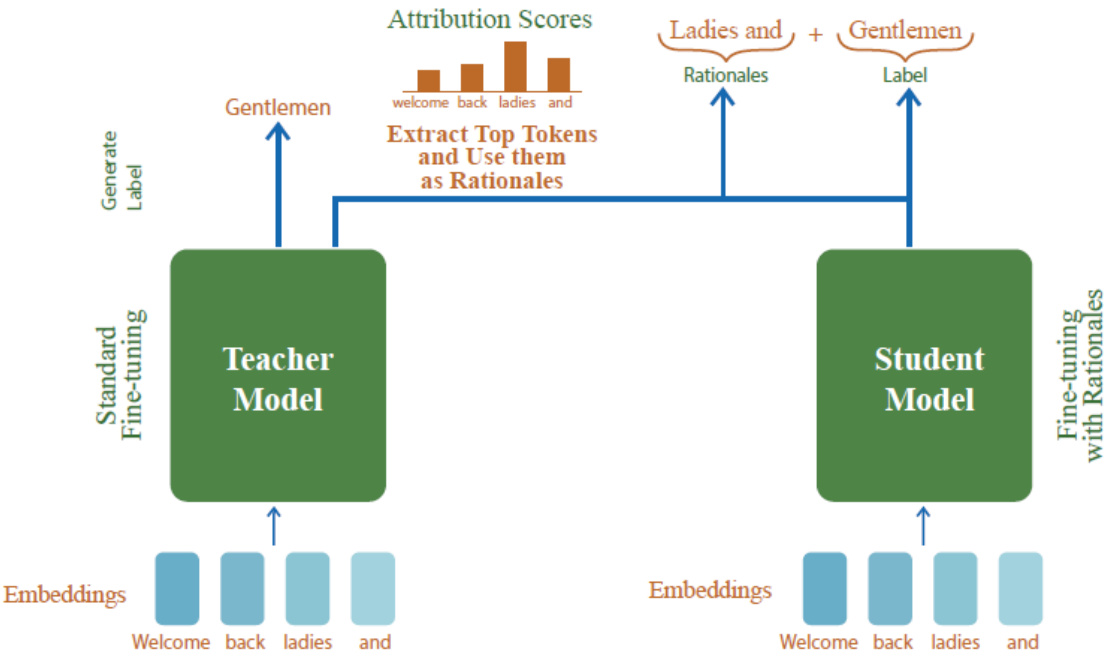
  - Datenschutzrechtliche Probleme werden sich auch in Zukunft in Europa stellen.

  - **Für den akademischen Bereich könnten kleine, aber spezialisierte Modelle relevant werden, die auch selbst trainiert und gehostet werden könnten**
  - **→ Knowledge Distillation**

Knowledge Distillation

- Idee: Kleine Modelle („Student Model“) werden von einem großen Modell („Teacher Model“) (semi-) automatisch trainiert.

- Möglichkeit 1 („White-Box Model“)
  - Relevanteste Token für den Output werden an das „Student Model“ weitergegeben
  - Dies verbessert das Lernen des „Student Models“ signifikant



Ballout et al. (2024a)

Knowledge Distillation

- Idee: Kleine Modelle („Student Model“) werden von einem großen Modell („Teacher Model“) (semi-) automatisch trainiert.

- Möglichkeit 2 („Black-Box Model“)
  - Step-by-step Erklärungen von einem großen Modell werden an das „Student Model“ weitergegeben.
  - Signifikante Verbesserung des Lernens
  - „Chain-of-thought“

Input:	[MIN 5 20 [MAX 12 26 [MED 1 23 [SUM 15 27 1]]]
Output: Without Explanation	5
Output: Short Explanation	SUM, 3 MED, 3 Max, 26 Min, 5 Final answer: 5
Output: Medium Explanation	SUM, add [15 27 1], sum is 43, answer is 3 MED, sorted list, [1 3 23], MEDIAN is 3 Max, largest number of [12, 26, 3] is 26 Min, smallest number of [5, 20, 26] is 5, Final answer: 5
Output: Long Explanation	SUM operator so we have to add the numbers [15 27 1], sum is 43 since it is modular sum, we choose the last number of the answer which is 3 MEDIAN operator so we sort the sequence first [1 3 23], MEDIAN is 3 Maximum operator so we choose the maximum number of sub-sequence [12, 26, 3] which is 26 Minimum operator so we choose the minimum number of sub-sequence [5, 20, 26] which is 5 Final answer: 5

FIGURE 4.2: Figure shows the three type of explanations produced for a short sample of the input

Ballout et al. (2023a)

Knowledge Distillation

- Idee: Kleine Modelle („Student Model“) werden von einem großen Modell („Teacher Model“) (semi-) automatisch trainiert.

- Möglichkeit 3 („Black-Box Model“)
  - ChatGPT beantwortet nicht nur die Task, sondern erklärt auch seine Antwort.
  - Z.B. waren für ein bestimmtes Dateset die 5 W-Fragen relevant (Who, What, Where, When, Why)
  - „Rationales“ werden von ChatGPT zur Verfügung gestellt.

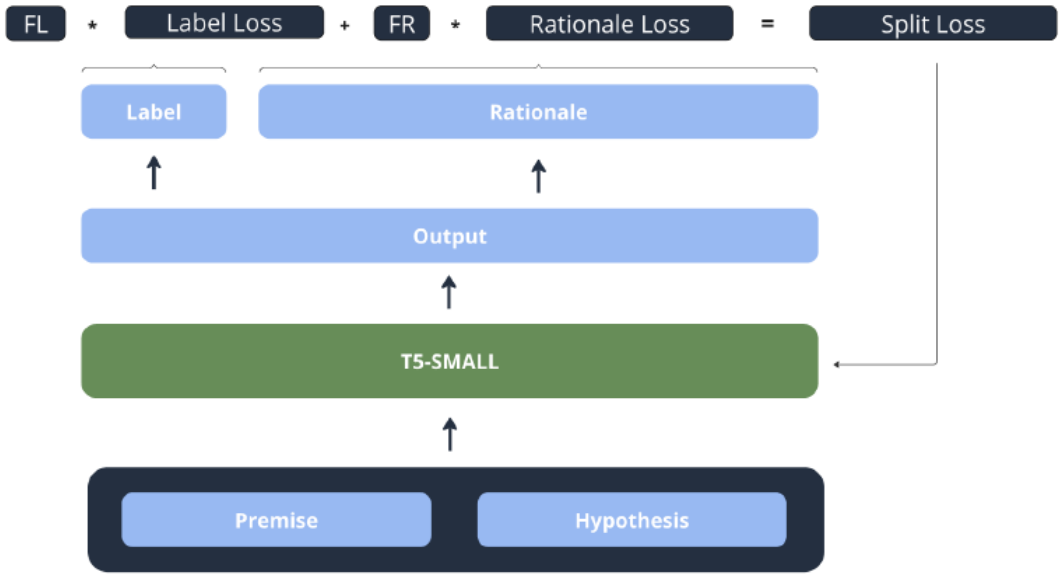
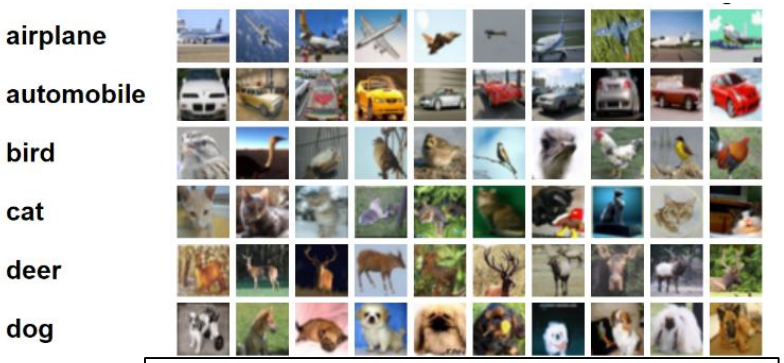


FIGURE 4.10: Fine-Tuning process of T5-Small.

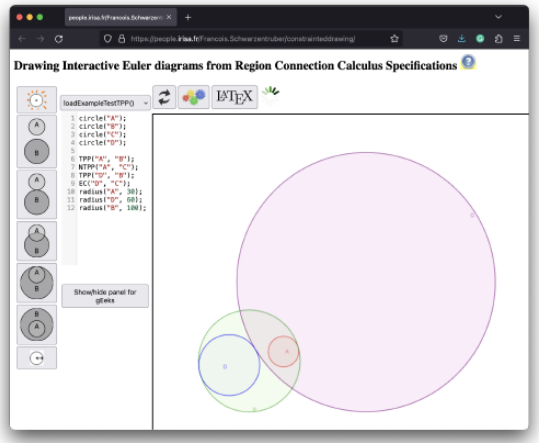
Pieper et al. (2024)

Transfer von Wissen

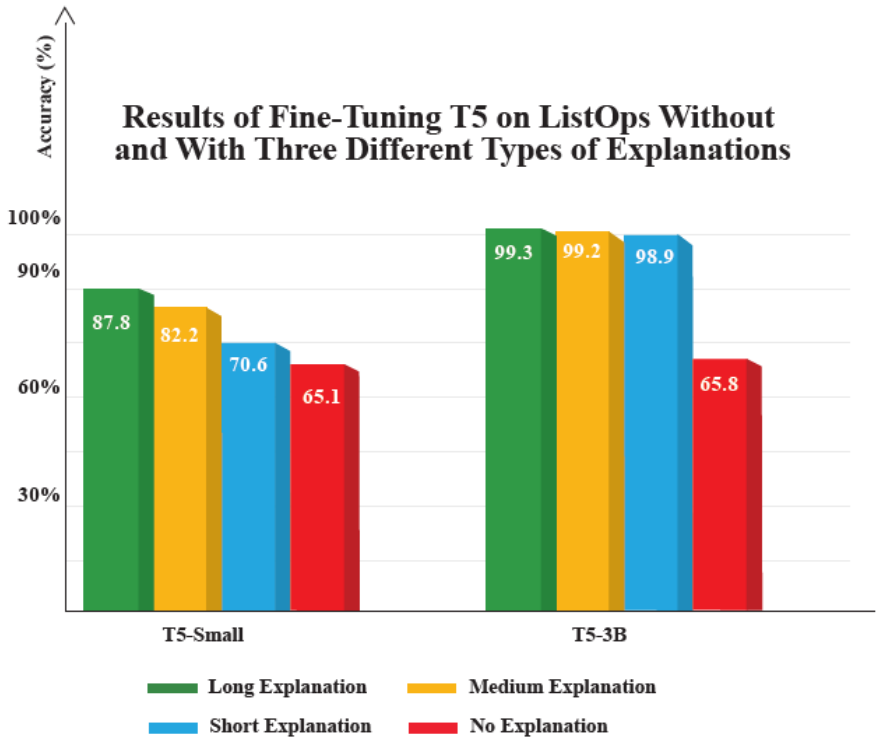
- Vortrainierte Sprachmodelle können durch Fine-Tuning signifikant effizienter und besser für völlig andere Domänen nachtrainiert werden können -> Transfer Learning
  - Bilderkennung, Beschreibung von Diagrammen, formale Sprache etc.



https://www.cs.toronto.edu/~kriz/cifar.html



Schorlemmer et al. (2024)



Ballout et al. (2023b)

Möglichkeiten der Anwendung von KI-Technologie

Welche konkreten Anwendungen sind vorstellbar?

Anwendungen im Mittelstand

- Beispielhafte Anwendungsmöglichkeiten
  - Viele Büroaufgaben könnten von KI-Systemen übernommen werden

  - Dienstleistungen im Wissensbereich können von Transformern / KI-Systemen unterstützt werden
    - Wirtschaftsprüfung & Steuer- und Unternehmensberatung, Reporting-Bedarfe
    - Unterstützung im Bereich des Endkunden

  - Datenanalyse
    - Nicht-linear Relationen zwischen Daten zu finden ist eine Stärke des maschinellen Lernens

  - Bildung als lebenslanges Lernen wird völlig neu gedacht werden.

Büroaufgaben

- Was man heute schon machen kann
  - Vorformulierte Mails
  - Vorsortierung und Klassifikation von Dokumenten
  - Zusammenfassung von längeren Dokumenten
  - Automatisierte (Vor-)Bewertung umfangreicher Statistiken
  - Teilautomatisierte Erstellung von Berichten

Dienstleistungen

- Dienstleistungen
  - Dienstleistungen im Personalbereich (HR):
    - Effizienteres Screening, bessere Planung, besserer Einsatz von Ressourcen, weniger Fehler…
    - KI kann zur Kostenreduktion und zur Qualitätssteigerung im HR-Bereich beitragen.

  - Können wir uns Beratungsdienstleistungen in Zukunft noch ohne KI vorstellen?
    - Finance, Wirtschaftsprüfungen, Steuerberatung, Unternehmensberatung, Unterstützung im Bereich des Endkunden, Verkaufsberatung etc.

  - Juristische Dienstleistungen im Bereich Vertragsrecht, rechtliche Beratung, Patente / IP etc.

Dienstleistungen

Stellen Sie sich vor, Sie arbeiten für eine Wirtschaftsprüfung und sollen ein betriebswirtschaftliches Problem bewerten:

Input:

Für die Durchführung der Maschinenbelegungsplanung liegen die Auftragsmengen (Stückzahlen) für fünf Produkte mit unterschiedlichen variablen Kosten pro Maschinenminute (Grenzkostensätze) und unterschiedlichen Bearbeitungszeiten vor. Die Fertigungskapazitäten der beiden Maschinen sind begrenzt.

Die konkreten Daten finden Sie in nachfolgender Tabelle:

Produkt	Auftrags-menge	Maschine A		Maschine B	
		min./Stck.	EUR/Stck.	min./Stck.	EUR/Stck.
1	250	6	1,20	6	1,50
2	500	9	1,80	5	1,25
3	400	6	1,20	4	1,00
4	1.000	5	1,00	3	0,75
5	200	4	0,80	4	1,00
Kapazität		6.000 Minuten		6.000 Minuten	
Grenzkostensatz		0,20 EUR/min.		0,25 EUR/min.	

a) Nehmen Sie zunächst an, es existieren keine Kapazitätsgrenzen. Welches Produkt sollte auf welcher Maschine gefertigt werden, damit die gesamten variablen Kosten des Maschineneinsatzes minimiert werden? Begründen Sie Ihre Antwort kurz! Wie hoch sind die gesamten Kosten des Maschineneinsatzes? **9 Punkte**

- KI-Output:

Produkt	Auftrags-menge	Maschine A min/Stck.	EUR/Stck.	Maschine B min/Stck.	EUR/Stck.
1	250	6	1,20	6	1,50
2	500	9	1,80	5	1,25
3	400	6	1,20	4	1,00
4	1.000	5	1,00	3	0,75
5	200	4	0,80	4	1,00

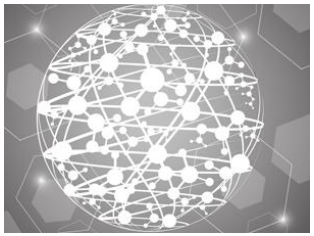
Schritte zur Berechnung:
1.Berechnung der variablen Kosten pro Produkt und Maschine:
…
…

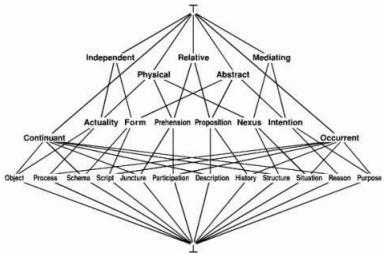
Fazit:
Um die variablen Kosten des Maschineneinsatzes zu minimieren, sollten die Produkte wie folgt gefertigt werden:
- Produkt 1 auf Maschine A
- Produkt 2 auf Maschine B
- Produkt 3 auf Maschine B
- Produkt 4 auf Maschine B
- Produkt 5 auf Maschine A

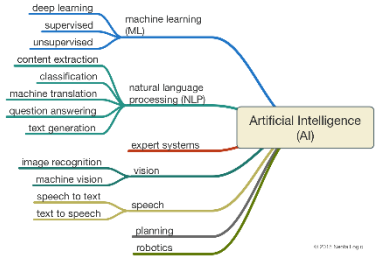
Die gesamten Kosten des Maschineneinsatzes betragen 2.235 EUR.

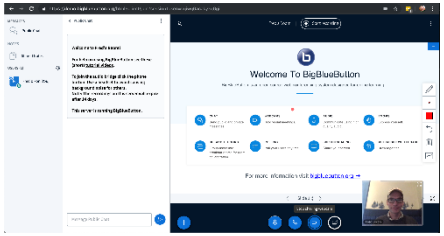
Bildung

- Bildung
  - Relevante Gruppen: Schüler, Studierende, Auszubildende, Berufstätige, Silver Ager, etc.
  - Lebenslanges Lernen wird das Prinzip der Zukunft
  - Viele Lernformate werden von digitalen Assistenten unterstützt werden.
  - Lernende müssen sich neue Schlüsselqualifikationen aneignen.
  - KI-Systeme werden Lernende früher oder später auch autonom bewerten.









Statt einer Zusammenfassung

- Wird es eine Superintelligenz / Singularität in absehbarer Zeit geben?
  - **NEIN**
- Wird man in Zukunft immer verstehen was eine KI macht?
  - **NEIN**
- Werden wir in Zukunft stärker durch KI überwacht und manipuliert werden?
  - **JA**
- Werden wir uns schnell genug an KI anpassen?
  - **HOFFENTLICH**
- Werden durch KI neue Chancen für Unternehmen entstehen
  - **JA**

Statt einer Zusammenfassung

- Werden neue soziale Strukturen entstehen?
  - **JA („Society of Minds and Machines“)**
- Werden rechtliche Rahmenbedingungen für den KI-Einsatz etabliert werden?
  - **HOFFENTLICH**
- Wird KI in militärischen und sicherheitsrelevanten Kontexten in Zukunft eine größere Rolle spielen?
  - **JA**
- Werden die Vorteile der KI die Nachteile und Gefahren überwiegen?
  - **JA**
- Wird sich die Bildung fundamental verändern?
  - **JA**

Vielen Dank !!!

Fragen, Kommentare, Anmerkungen….