Crash Kurs KI

„Kritische KI-Kompetenz sollte integraler Bestandteil wissenschaftlicher Integrität und philosophischer Bildung sein.“

Leitlinien zum Umgang mit generativer KI am Institut für Philosophie der FU Berlin

Montag 22.6.2026 12:15 Uhr	***Grundlagen*** *Symbolische vs. subsymbolisch KI, Künstliche neuronale Netze, Interpretierbarkeit (Vortrag: Miguel Hoeltje)*
Montag 29.6.2026 12:15 Uhr	***Large Language Models und Chatbots*** *Lern-Paradigmen, Transformer-Architektur, LLMs vs. Chatbots, Reasoner-Models, Multimodalität (Vortrag: Miguel Hoeltje)*
Montag 6.7.2026 12:15 Uhr	***Über Large Language Models hinaus*** *Agentische KI, Vision-Language-Action Modelle, „Bitter Lesson“ & „Era of Experience“, World Models, Neurosymbolische KI (Vortrag: Miguel Hoeltje)*
Montag 13.7.2026 12:15 Uhr	***Zur ethischen Dimension der Entwicklung und Verwendung von KI*** *(Vortrag: Luise Müller)*

Folien, Handouts, ggfs. aktuelle Infos, etc.?

www.crash-ai.de

Der Plan für diese Sitzung

1. Rekap (& Ergänzung): Transformer-basierte *Large Language Models* am Beispiel GPT-3

2. Bewertung von *Large Language Models*: Stärken & Schwächen, Erfolge & Kritik

3. Beispiel für Weiterentwicklungen: *Vision-Language-Action Models* (VLAs)

4. Beispiel für Weiterentwicklungen: *Joint Embedding Predictive Architecture* (JEPA)

Ich war bisher zu langsam; aber es scheint mir wichtig, auf die Transformer noch mal einzugehen…

Nur eine ***kleine Auswahl*** an „Erfolgen“ & KI-interner Kritik

Falls noch Zeit ist und auch dann nur sehr skizzenhaft

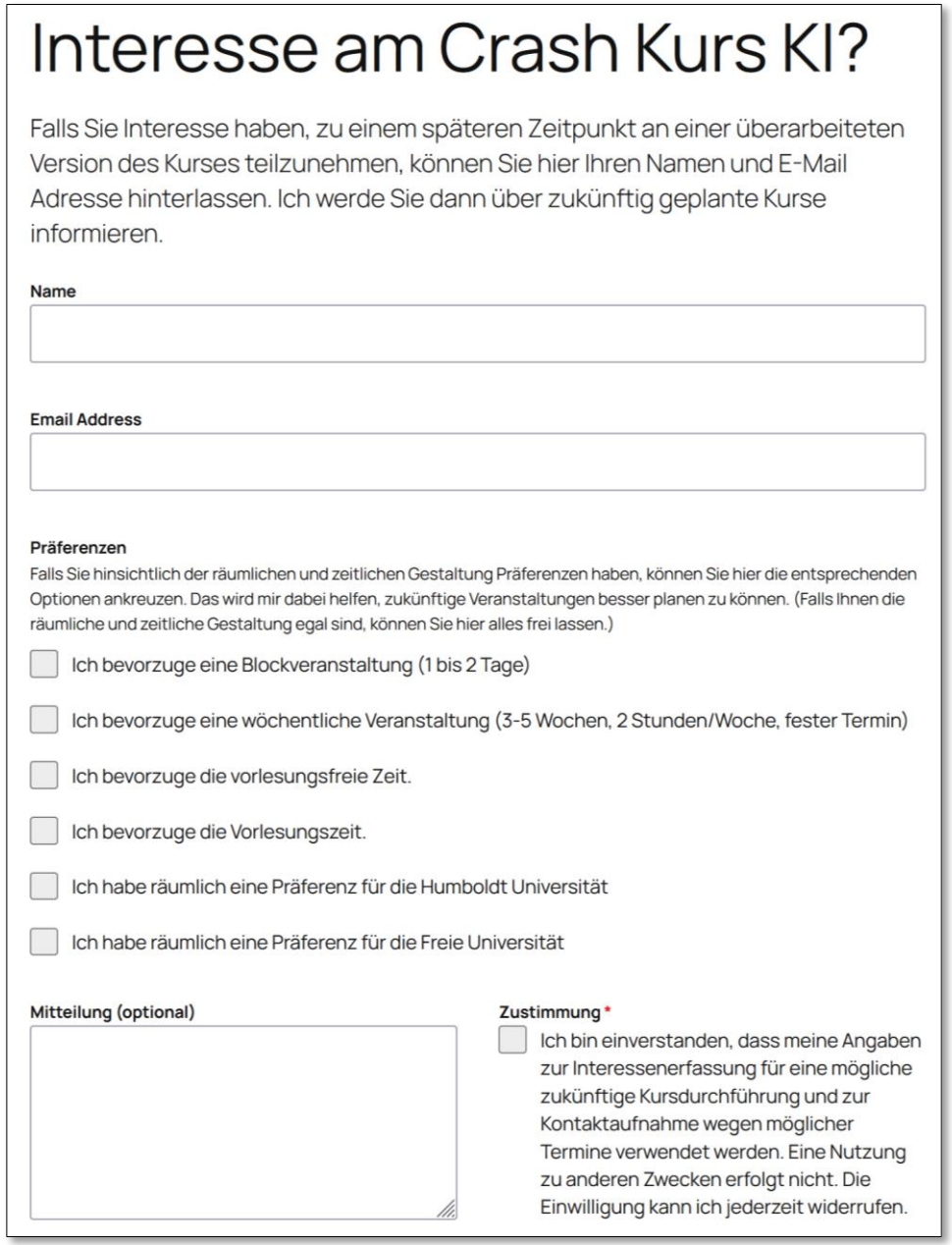
Auch hier: ***Nur 2 Beispiele*** aus einem riesigen Feld von aktiven Forschungs-projekten

Zukünftige Crash-Kurse

Falls Sie Interesse an einer überarbeiteten/erweiterten Version des Crash-Kurse haben, können Sie sich auf

www.crash-ai.de

jetzt in die Mailingliste eintragen.



Interesse am Crash Kurs KI?

Falls Sie Interesse haben, zu einem späteren Zeitpunkt an einer überarbeiteten Version des Kurses teilzunehmen, können Sie hier Ihren Namen und E-Mail Adresse hinterlassen. Ich werde Sie dann über zukünftig geplante Kurse informieren.

Name

Email Address

Präferenzen

Falls Sie hinsichtlich der räumlichen und zeitlichen Gestaltung Präferenzen haben, können Sie hier die entsprechenden Optionen ankreuzen. Das wird mir dabei helfen, zukünftige Veranstaltungen besser planen zu können. (Falls Ihnen die räumliche und zeitliche Gestaltung egal sind, können Sie hier alles frei lassen.)

- ☐ Ich bevorzuge eine Blockveranstaltung (1 bis 2 Tage)
- ☐ Ich bevorzuge eine wöchentliche Veranstaltung (3-5 Wochen, 2 Stunden/Woche, fester Termin)
- ☐ Ich bevorzuge die vorlesungsfreie Zeit.
- ☐ Ich bevorzuge die Vorlesungszeit.
- ☐ Ich habe räumlich eine Präferenz für die Humboldt Universität
- ☐ Ich habe räumlich eine Präferenz für die Freie Universität

Mitteilung (optional)

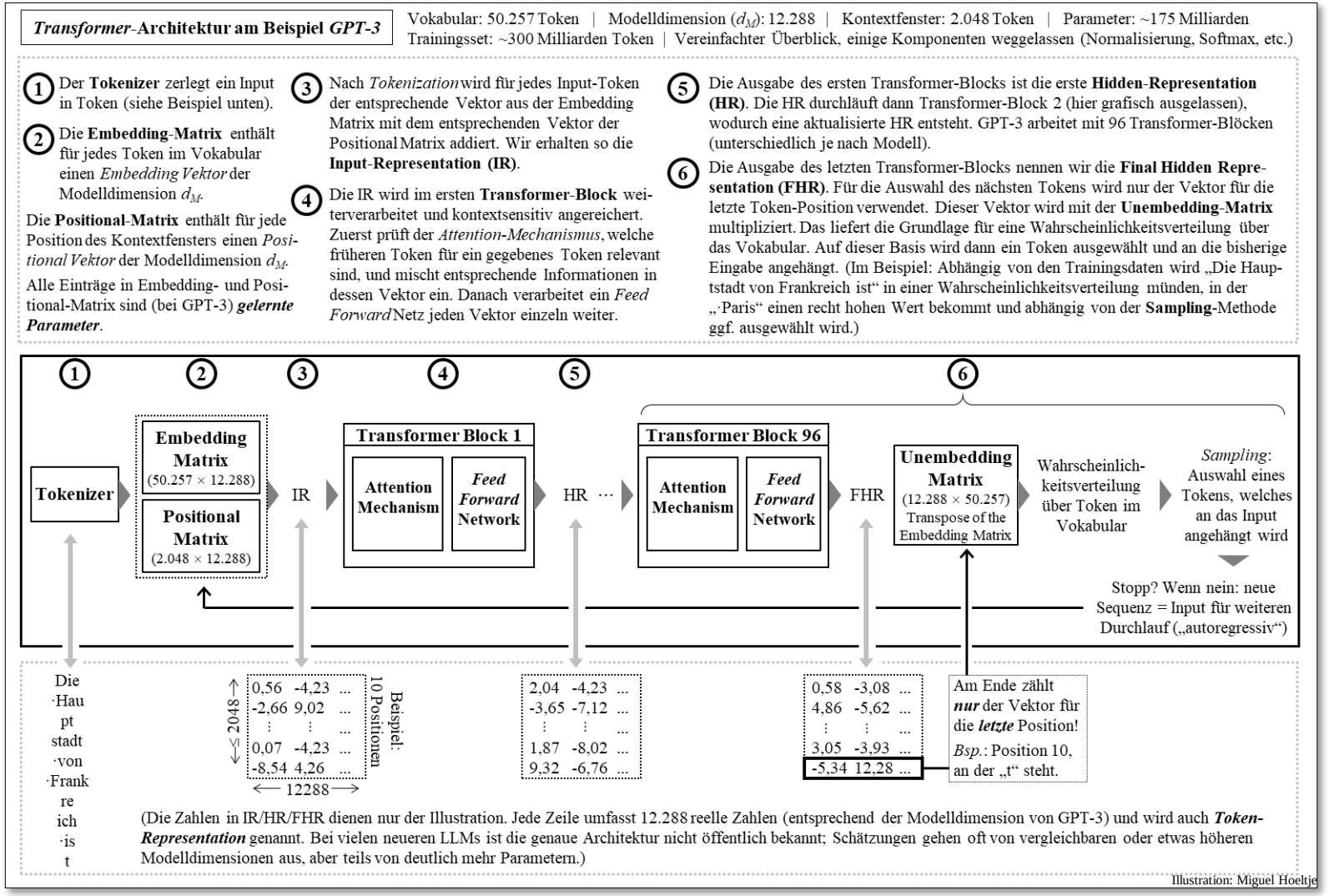
Zustimmung *

☐ Ich bin einverstanden, dass meine Angaben zur Interessenerfassung für eine mögliche zukünftige Kursdurchführung und zur Kontaktaufnahme wegen möglicher Termine verwendet werden. Eine Nutzung zu anderen Zwecken erfolgt nicht. Die Einwilligung kann ich jederzeit widerrufen.

Transformer LLMs | *Handout*

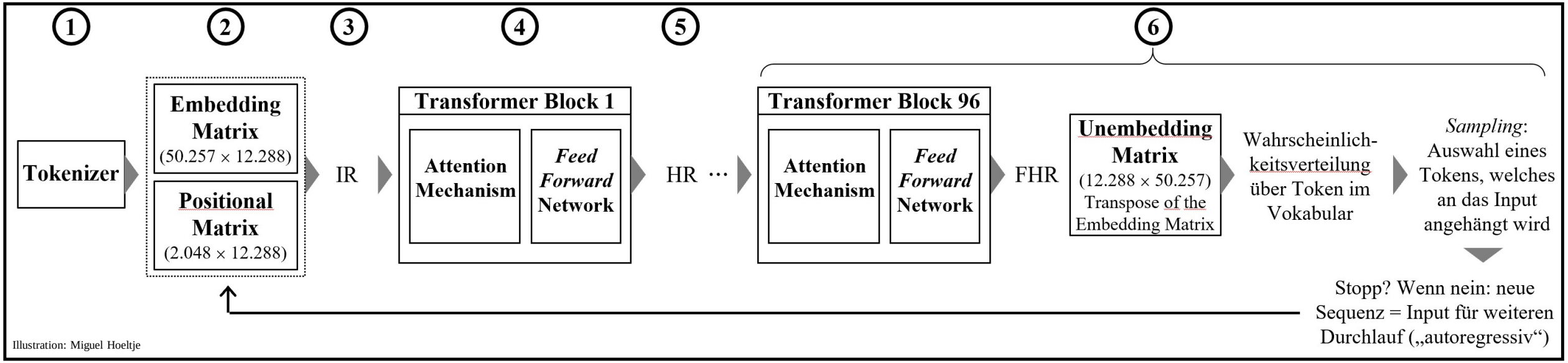
Handout: **_Transformer_-Architektur am Beispiel _GPT-3_**

Um die Komponenten eines Transformer-LLMs kennenzulernen, gehen wir dieses Handout schrittweise durch.



Transformer LLMs | *Übersicht*

Gesamtübersicht (Vereinfachter Überblick, einige Komponenten weggelassen (Normalisierung, Softmax, etc.).



Transformer LLMs | *Input ⇨ IR*

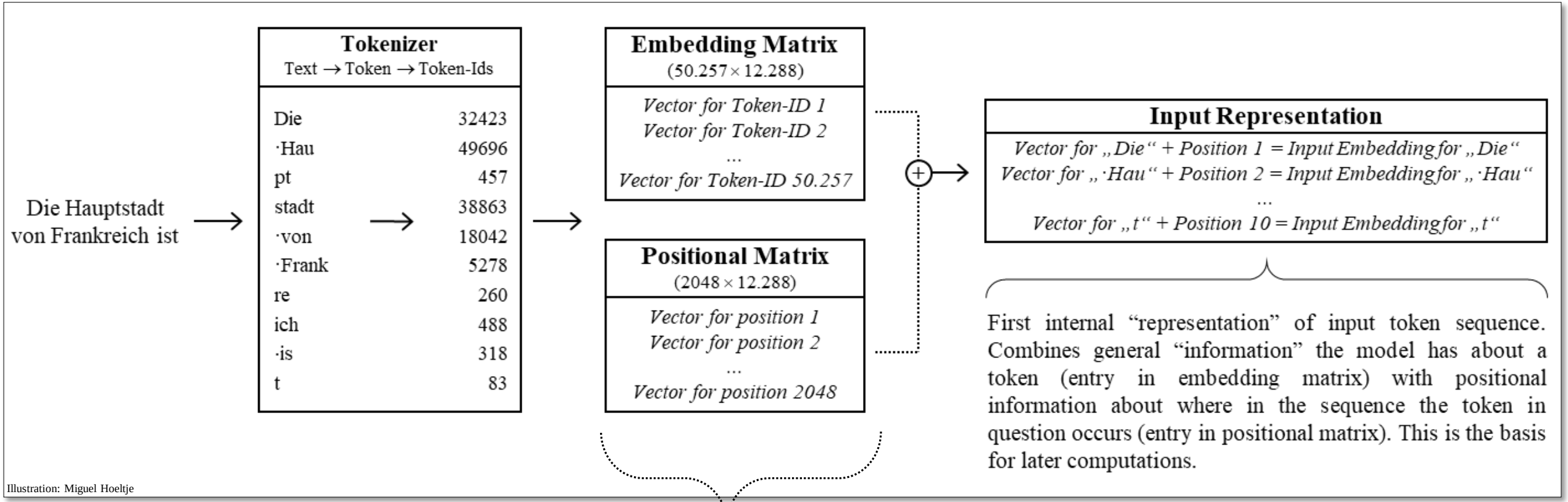


Illustration: Miguel Hoeltje

Die **Embedding-Matrix** enthält für jedes Token im Vokabular einen *Embedding Vektor* der Modelldimension d_M.

Die **Positional-Matrix** enthält für jede Position des Kontextfensters einen *Positional Vektor* der Modelldimension d_M.

Alle Einträge in Embedding- und Positional-Matrix sind (bei GPT-3) ***gelernte Parameter***.

Transformer LLMs | *Input ⇨ IR*

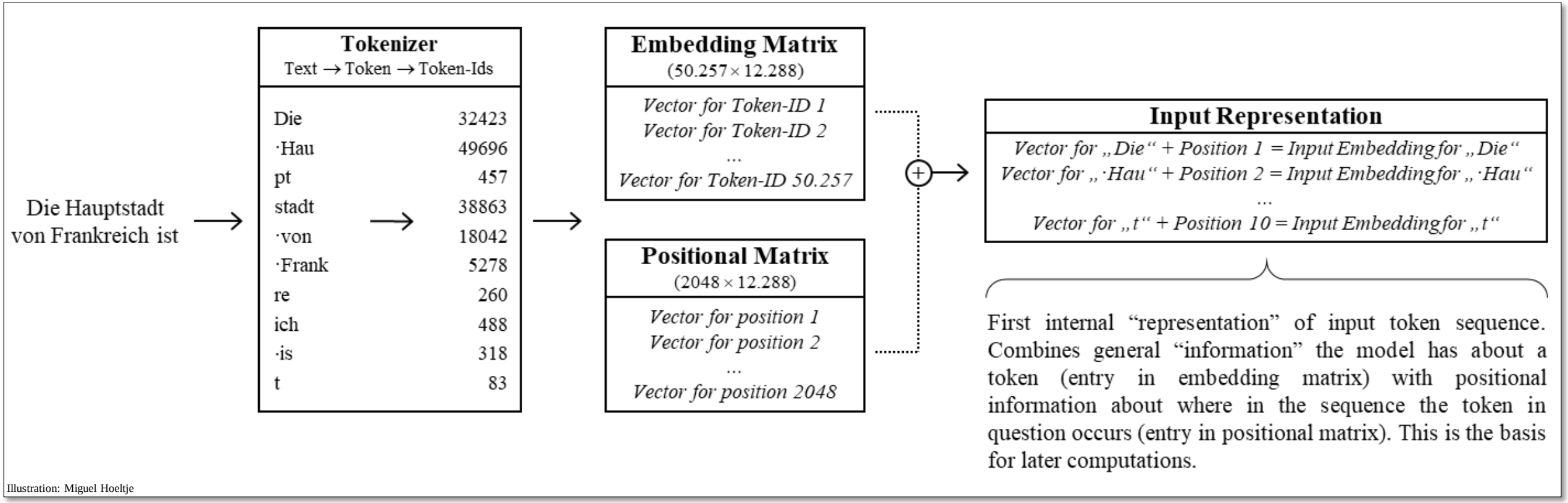
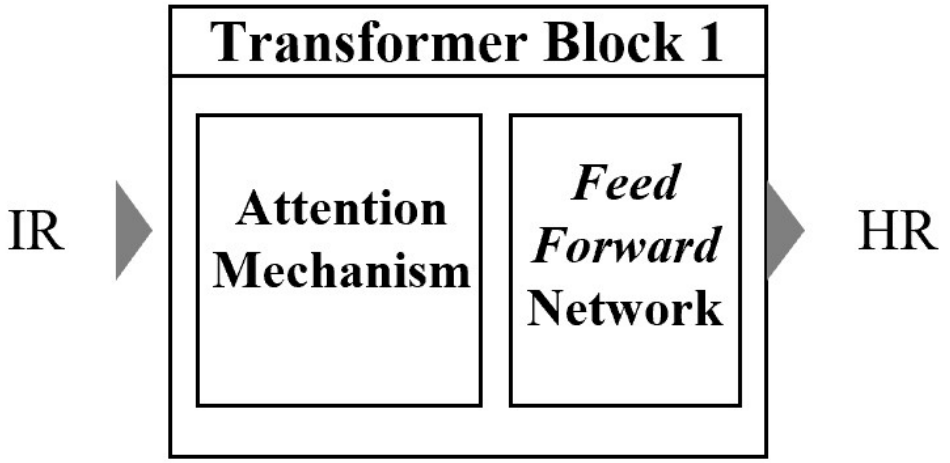


Illustration: Miguel Hoeltje

Nach *Tokenization* wird für jedes Input-Token der entsprechende Vektor aus der Embedding Matrix mit dem entsprechenden Vektor der Positional Matrix addiert. Wir erhalten so die **Input-Representation (IR)**.

Transformer LLMs | *Transformer Blocks*

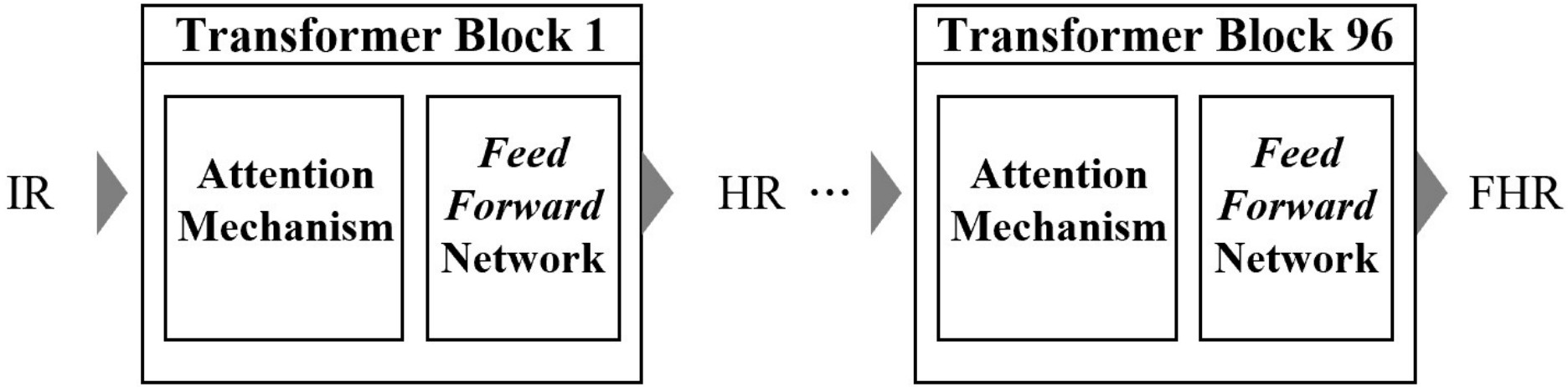


Die IR wird im ersten **Transformer-Block** weiterverarbeitet und kontextsensitiv angereichert.

Zuerst prüft der ***Attention-Mechanismus***, welche früheren Token für ein gegebenes Token relevant sind, und mischt entsprechende Informationen in dessen Vektor ein.

Danach verarbeitet ein ***Feed Forward* Netz** jeden Vektor einzeln weiter.

Transformer LLMs | *Transformer Blocks*



GPT-3 umfasst 96 *Transformer Blocks*…

Transformer LLMs | *Transformer Blocks*

IR/HR
(*n* x 12288)

→

Attention Layer

GPT-3: 96 Attention-Heads, die parallel arbeiten. Wir ignorieren zur Vereinfachung die Aufsplittung in einzelne Heads.

W_Q Query Matrix

W_K Key Matrix

W_V Value Matrix

Jede dieser Matrizen hat die Größe 12288 x 12288, enthält also rund 151 Millionen gelernte Parameter.

Durch Matrix-Multiplikation mit IR/HR erhalten wir für jede *Token-Representation* I:

- Query-Vektor (wonach sucht I?)
- Key-Vektor (worauf antwortet I?)
- Value-Vektor (was steuert I bei?)

Für jede Token-Representation I:

- Vergleiche den *Query*-Vektor mit den *Key*-Vektoren der *Token-Representations*, die vor *I* (bis inklusive *I*) kommen.
- Dies bildet die Grundlage für eine Gewichtung, wie relevant frühere *Token-Representations* für *I* sind.
- Dies nutzen wir, um eine gewichtete Summe der *Value*-Vektoren der anderen *Token-Representations* zu bilden.
- Diese Summe leistet dann einen Beitrag zum Update der *Token-Representation I*.

Illustration: Miguel Hoeltje

↓

Aktualisierte HR (+ Residual Stream)

↓

Transformer LLMs | *Transformer Blocks*

Feed Forward Network

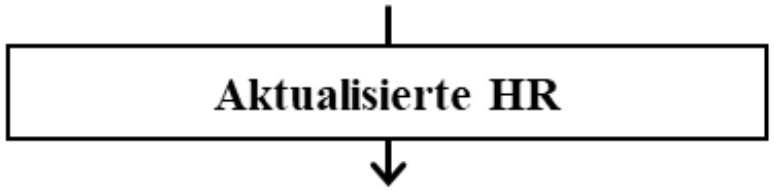
(Häufig auch *Multi Layer Perceptron* (MLP) genannt)

Ein einfaches neuronales Netz (strukturell analog zu dem Beispiel, welches wir uns in der letzten Sitzung für die Ziffern-Erkennung angeschaut hatten).

Alle *Token-Representations* durchlaufen das MLP einzeln, es gibt keine „Interaktion" zwischen den *Representations* für einzelne Token (anders als in der Attention Layer, wo *Token-Representations* über den Query/Key/Value-Mechanismus andere *Token-Representations* beeinflussen können).

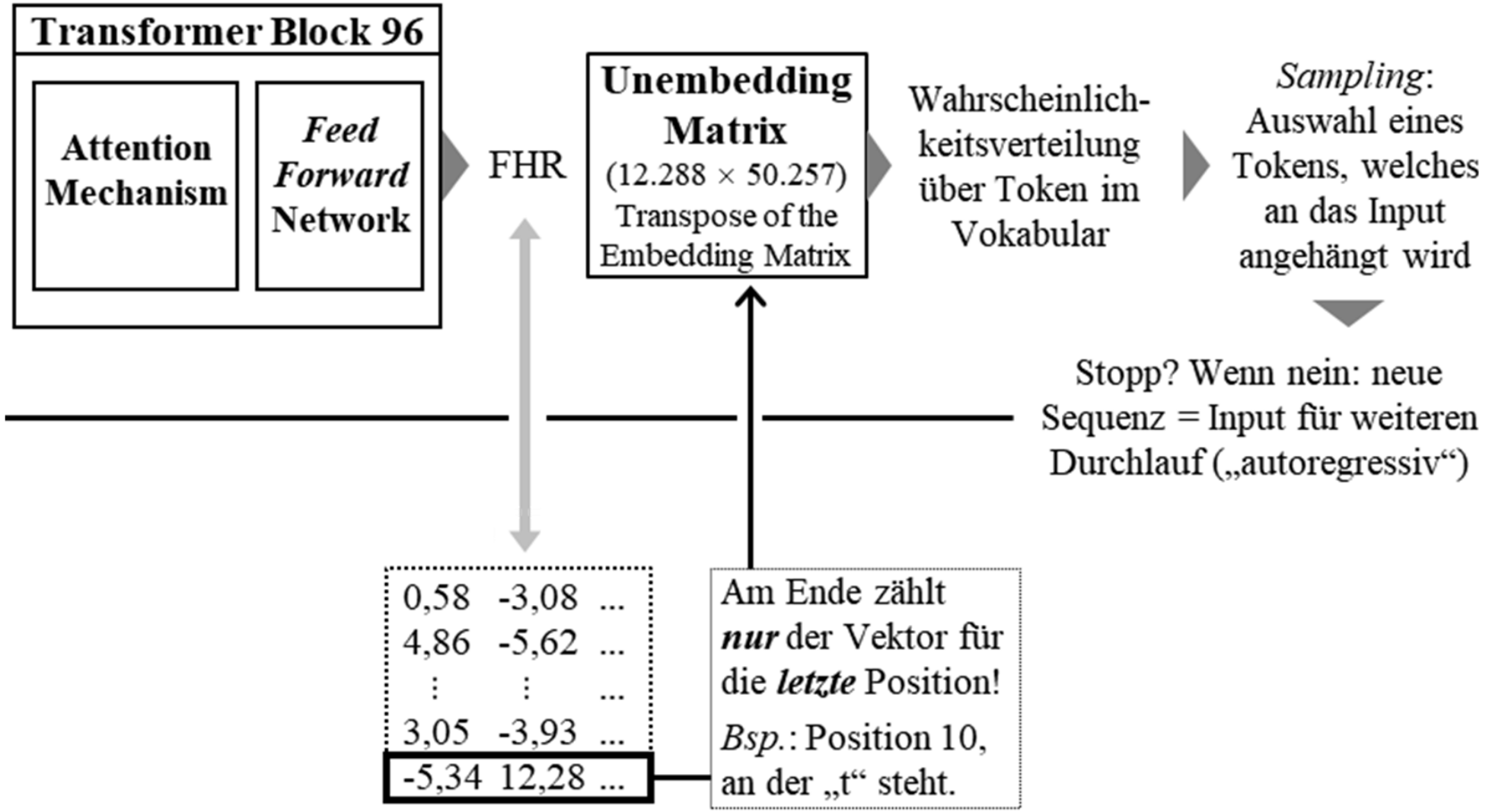
Die typische Rolle des MLP ist: die gerade durch Attention kontextualisierte *Token-Representations* nichtlinear „umzuformen"—Features verstärken/unterdrücken, neue Merkmale kombinieren, Bedeutungsaspekte umkodieren. Es ist der Teil, der pro Token „Rechenarbeit" macht, während Attention die Informationen „zusammenträgt".

Jedes der MLP bei GPT-3 umfasst allein 1,21 Milliarden Parameter.



(+ Residual Stream)

Transformer LLMs | *FHR ⇨ Output*



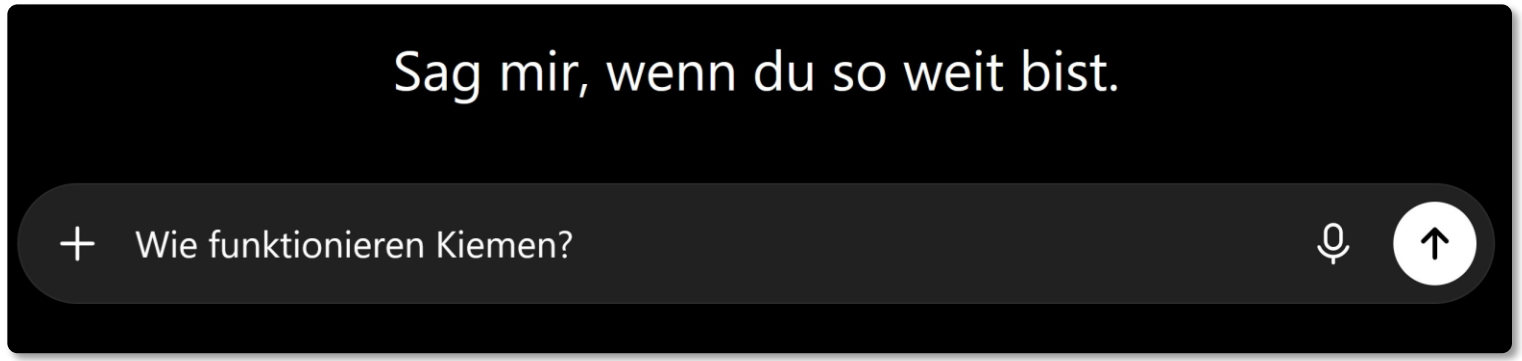
GPT vs. ChatGPT

Wir müssen ein reines LLM (das neuronale Netz) unterscheiden von dem größeren Gesamtsystem, in welches das LLM eingebettet ist.

GPT	Das LLM
ChatGPT	Das Gesamtsystem, mit dem User interagieren

Während das LLM Text entgegennimmt und autoregressiv (Token-für-Token) einen Output generiert, verfügt dass Gesamtsystem über weitere Komponenten.

User-Prompt *vs.* vollständiges Input



Wenn wir als User mit einem LLM interagieren, bildet unser Prompt in der Regel nur einen kleinen Teil des tatsächlichen Inputs, welches das LLM entgegen nimmt.

Der Rest ist für uns als User nicht sichtbar.

Das vollständige Input umfasst zusätzlich zum Prompt in der Regel noch:

System message + developer instructions
(general rules, safety rules, …)

User profile

Conversation context

Current user prompt

External material

Das vollständige Input kann selbst bei kurzen Prompts mitunter sehr lang werden …

Autoregression | *Token-für-Token…*

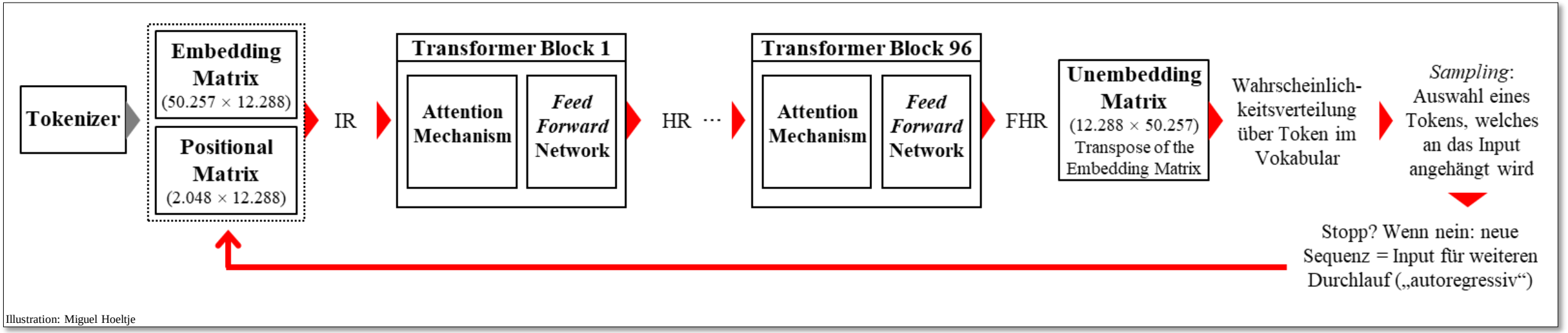


Illustration: Miguel Hoeltje

Autoregressiv heißt:

Input I ⇨ LLM ⇨ ein neues Token T ⇨ I+T = neuer Input ⇨ LLM ⇨ ein neues Token T* ⇨ …

deshalb auch extrem rechenintensiv (vor allem, wenn man noch *System Prompt* + *Conversation Context* + *External Material* + … mitbedenkt).

(Und damit extrem und zunehmend teuer ⇨ bislang keine wirklichen Profite bei *OpenAI*, *Anthropic*, …)

Autoregression | *Token-für-Token…*

An **autoregressive model** (or AR model) is one in which each element x_i of the data vector $\mathbf{x}$ is predicted based on other elements of the vector. Such a model has no latent variables. If $\mathbf{x}$ is of fixed size, an AR model can be thought of as a fully observable and possibly fully connected Bayes net. This means that calculating the likelihood of a given data vector according to an AR model is trivial; the same holds for predicting the value of a single missing variable given all the others, and for sampling a data vector from the model.

The most common application of autoregressive models is in the analysis of time series data, where an AR model of order k predicts x_t given $x_{t-k}, \ldots, x_{t-1}$. In the terminology of Chapter 14, an AR model is a non-hidden Markov model. In the terminology of Chapter 24, an n-gram model of letter or word sequences is an AR model of order $n-1$.

In classical AR models, where the variables are real-valued, the conditional distribution $P(x_t \mid x_{t-k}, \ldots, x_{t-1})$ is a linear–Gaussian model with fixed variance whose mean is a weighted linear combination of $x_{t-k}, \ldots, x_{t-1}$—in other words, a standard linear regression model. The maximum likelihood solution is given by the **Yule–Walker equations**, which are closely related to the **normal equations** on page 698.

A **deep autoregressive model** is one in which the linear–Gaussian model is replaced by an arbitrary deep network with a suitable output layer depending on whether x_t is discrete or continuous. Recent applications of this autoregressive approach include DeepMind's WaveNet model for speech generation (van den Oord *et al.*, 2016a). WaveNet is trained on raw acoustic signals, sampled 16,000 times per second, and implements a nonlinear AR model of order 4800 with a multilayer convolutional structure. In tests it proves to be substantially more realistic than previous state-of-the-art speech generation systems.

Russell & Norvig (2022): *Artificial Intelligence: A Modern Approach*, 4th Edition, Global Edition

Autoregression | *Token-für-Token…*

Ein sehr gutes Video zur Funktionsweise und den Auswirkungen von Autoregression finden Sie hier:

Lern-Paradigmen

In **supervised learning** the agent observes input-output pairs and learns a function that maps from input to output. For example, the inputs could be camera images, each one accompanied by an output saying "bus" or "pedestrian," etc. An output like this is called a **label**. The agent learns a function that, when given a new image, predicts the appropriate label. In the case of braking actions (component 1 above), an input is the current state (speed and direction of the car, road condition), and an output is the distance it took to stop. In this case a set of output values can be obtained by the agent from its own percepts (after the fact); the environment is the teacher, and the agent learns a function that maps states to stopping distance.

In **reinforcement learning** the agent learns from a series of reinforcements: rewards and punishments. For example, at the end of a chess game the agent is told that it has won (a reward) or lost (a punishment). It is up to the agent to decide which of the actions prior to the reinforcement were most responsible for it, and to alter its actions to aim towards more rewards in the future.

Russell & Norvig (2022): *Artificial Intelligence: A Modern Approach*, 4th Edition, Global Edition

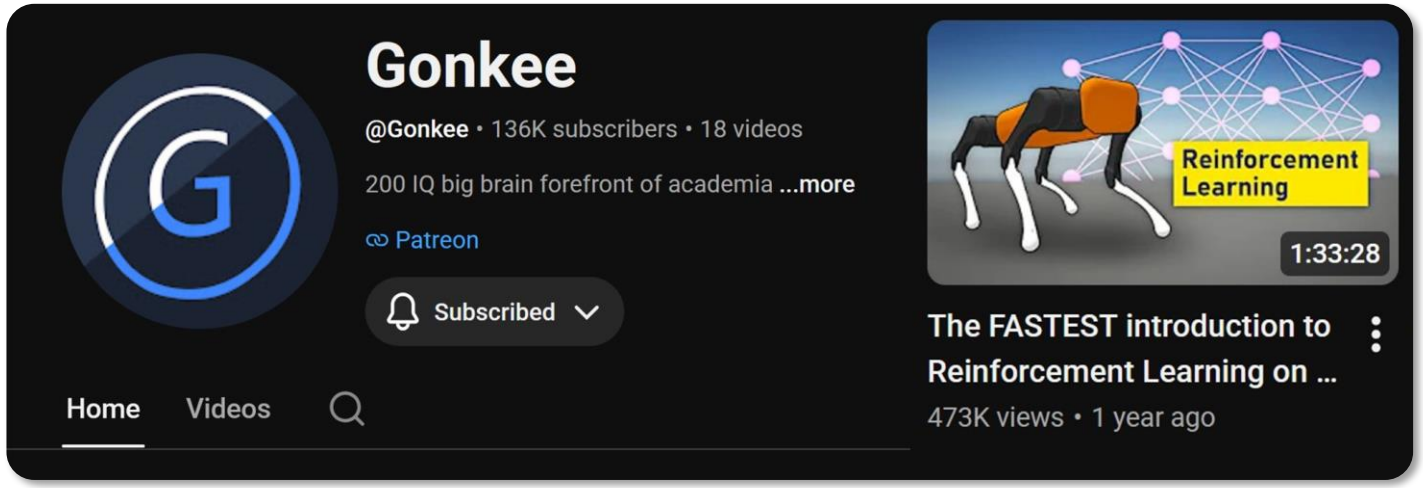
Lern-Paradigmen

Das ***Pre-Training*** von LLMs läuft mittels ***Self-Supervised-Learning*** (next token prediction, die Trainingsdaten bringen ihre eigenen Labels/Targets gleich mit)

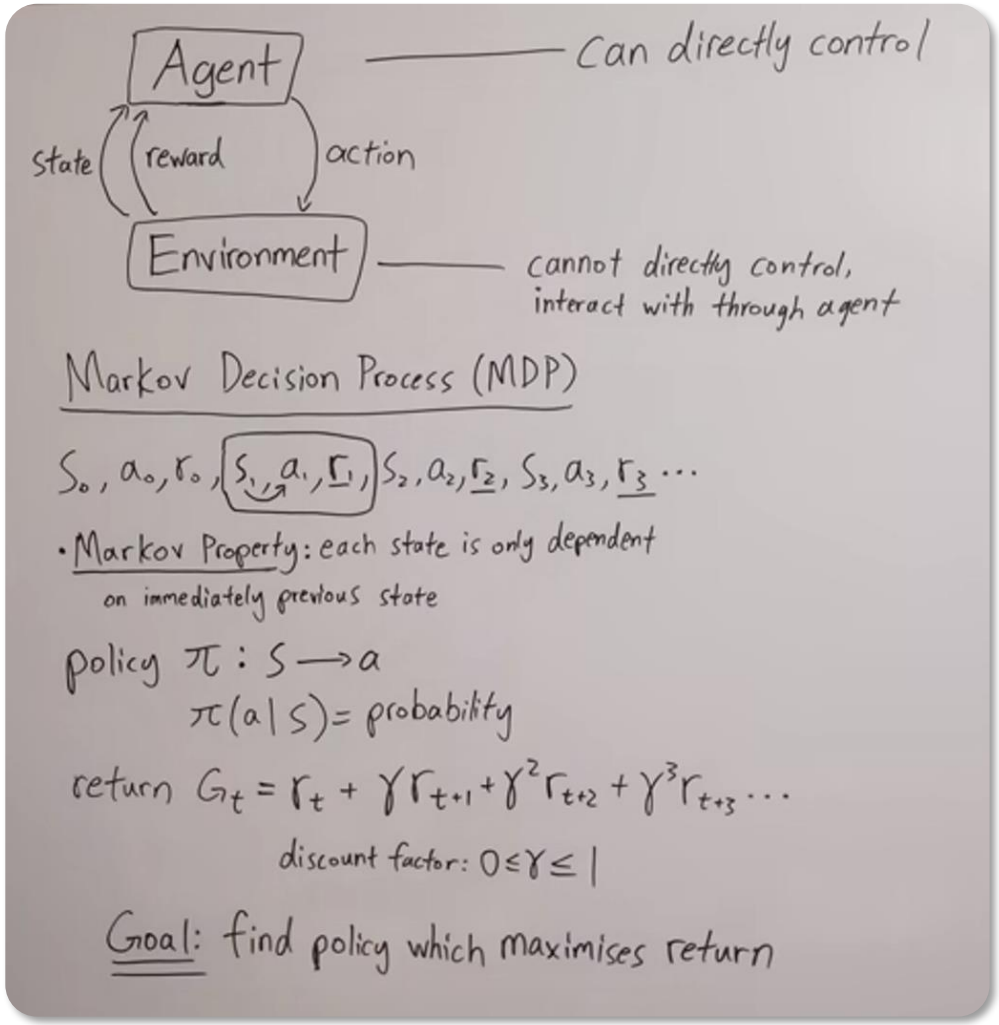
Nach dem Pre-Training wird in der Regel noch ***RLHF*** (reinforcement learning with human feedback) angewandt (oft unter Beteiligung weiterer KI-Systeme).

Reinforcement Learning

Ein sehr gutes Video zu *Reinforcement Learning* finden Sie hier:



https://www.youtube.com/watch?v=VnpRp7ZglfA



Wie sind wir bei LLMs gelandet?

Zusammenspiel mehrerer Faktoren:

- Transformer-Architektur (2017)
- Verfügbarkeit großer Mengen digitalen Textes für *Pre-Training*
- *Self-Supervised Learning* (Text bringt die eigenen Label/Targets automatisch mit)
- Zuwachs an Rechenleistung

LLMs: Erfolge

LLMs haben gezeigt:	Konversationsfähigkeit lässt sich erstaunlich gut auf der Basis von Vorhersagetraining „erlernen“ (mittlerweile vermutlich nahe am Bestehen des Turing-Tests …)
Grundprinzip	Sehr große Textmengen als Trainingsdaten, Transformer-Architektur, Training auf „nächstes Token vorhersagen“, danach Feintuning mit menschlichem Feedback.
Resultat	Flüssige Textproduktion, Übersetzung und Zusammenfassung, Code-Generierung, Few-Shot- und In-Context-Learning, teilweise überraschende Schlussfolgerungsfähigkeiten

LLMs: Erfolge

OpenAI-Erfolg in der Mathematik (2026)

- Ein internes OpenAI-Modell fand einen Beweis zu einem bekannten offenen Problem aus der Mathematik des 20. Jahrhunderts (es widerlegte eine lange diskutierte Vermutung von Paul Erdős).
- Der Beweis wurde anschließend von Mathematiker*innen geprüft. Tim Gowers bezeichnete das Ergebnis als wichtigen Meilenstein.
- Das Modell war kein speziell gebauter Theorembeweiser, sondern ein allgemeines *Reasoning-Modell* im Stil moderner LLMs. Nicht speziell für Mathematik trainiert, nicht auf dieses Problem zugeschnitten, nicht durch eine externe Proof-Search-Architektur geführt.
- Der Fall zeigt: LLMs können nicht nur Texte formulieren, sondern zu echter mathematischer Forschung beitragen.

OpenAI (2026): *Planar Point Sets with Many Unit Distances*. Technical proof write-up.

LLMs (und KI im weiteren Sinne): Erfolge

REMARKS ON THE DISPROOF OF THE UNIT DISTANCE CONJECTURE

NOGA ALON, THOMAS F. BLOOM, W. T. GOWERS, DANIEL LITT, WILL SAWIN, ARUL SHANKAR, JACOB TSIMERMAN, VICTOR WANG, AND MELANIE MATCHETT WOOD

Abstract. We present a short, digested, human-verified version of the recent OpenAI-generated counterexample to the Erdős unit distance conjecture, and a sequence of reflections on it. The argument relies crucially on ideas that may, at least in retrospect, be attributed to Ellenberg-Venkatesh, Golod-Shafarevich, and Hajir-Maire-Ramakrishna.

Alon, N. et al. (2026): *Remarks on the disproof of the unit distance conjecture*. arXiv:2605.20695.

LLMs (und KI im weiteren Sinne): Erfolge

4. Thomas Bloom

This was one of Erdős' favourite problems – he first asked it in 1946 [14] and returned to it many times. (The site www.erdosproblems.com, on which it is Problem #90, currently lists 14 separate references, and there are no doubt more.) The influential collection of 'Research Problems in Discrete Geometry' by Brass, Moser, and Pach [8] describes it as 'possibly the best known (and simplest to explain) problem in combinatorial geometry'. For an AI to produce a solution to a problem of this calibre is both surprising and impressive.

5. W T Gowers

Can we still identify some mathematical capability that human mathematicians have and AI does not yet have? If so, what might that capability be, and how could one go about demonstrating that AI still lacks it?

Almost certainly the answer to the first question will have to be quantitative rather than qualitative. That is, we are unlikely to be able to show that there is something we can do that current AI models cannot in principle do at all, but we might be able to show that there are things we can still do much more efficiently than those models. But when a model has just solved a major open problem, it is clear that even a modest conclusion like that will not be straightforward to demonstrate, and indeed isn't obviously true.

Alon, N. et al. (2026): *Remarks on the disproof of the unit distance conjecture*. arXiv:2605.20695.

Kritik an LLMs

Ich konzentriere mich hier auf „Kritik“ an LLMs aus der Perspektive der KI-Forschung:

Ziel (u.a.):	Entwicklung intelligenter Systeme
Mögliche Kritik:	Nicht der richtige Ansatz, zu sehr auf Text fokussiert, kein Lernen in Echtzeit, prinzipiell anfällig für Halluzinationen, etc.

Und auch hier greife ich nur ein paar Beispiele auf (es gibt viele andere…).

Politische/normative Kritik an der Entwicklung, dem Einsatz und der gesellschaftlichen Einbettung von LLMs (und anderer KI-Systeme) blende ich weitestgehend aus.

Aber: Nächste Woche!

Chomsky

There's a kind of a industry in computational cognitive science and computer science trying to show that you can get significant knowledge of a language by statistical analysis of text. Antecedently, that's extremely unlikely to succeed. You do not get discoveries in the sciences by taking huge amounts of data, throwing them into a computer, and doing statistical analysis of them. Try to think it through in the history of the sciences. It just doesn't happen. That's not the way you understand things. You have to have theoretical insights. You have to know what kind of experiments to carry out-- what kind of data are worth looking at, which kind of throw away, and so on. That's the way the sciences have always worked.

Noam Chomsky (2014): *Understanding Linguistics*, Talks at Google

Chomsky

If you wanted to, say, study the laws of motion, you could take a huge number of videotapes of what's happening outside the window, and subject them to statistical analysis. You could get a pretty good prediction of the next thing that's going to happen outside the window-- actually, a better prediction than what the physics department can give. But it's not science. It's a way of matching data, and maybe predicting some new data. But that's not what understanding is. And it's very unlikely to work for language, either.

Noam Chomsky (2014): *Understanding Linguistics*, Talks at Google

Richard Sutton

The Bitter Lesson

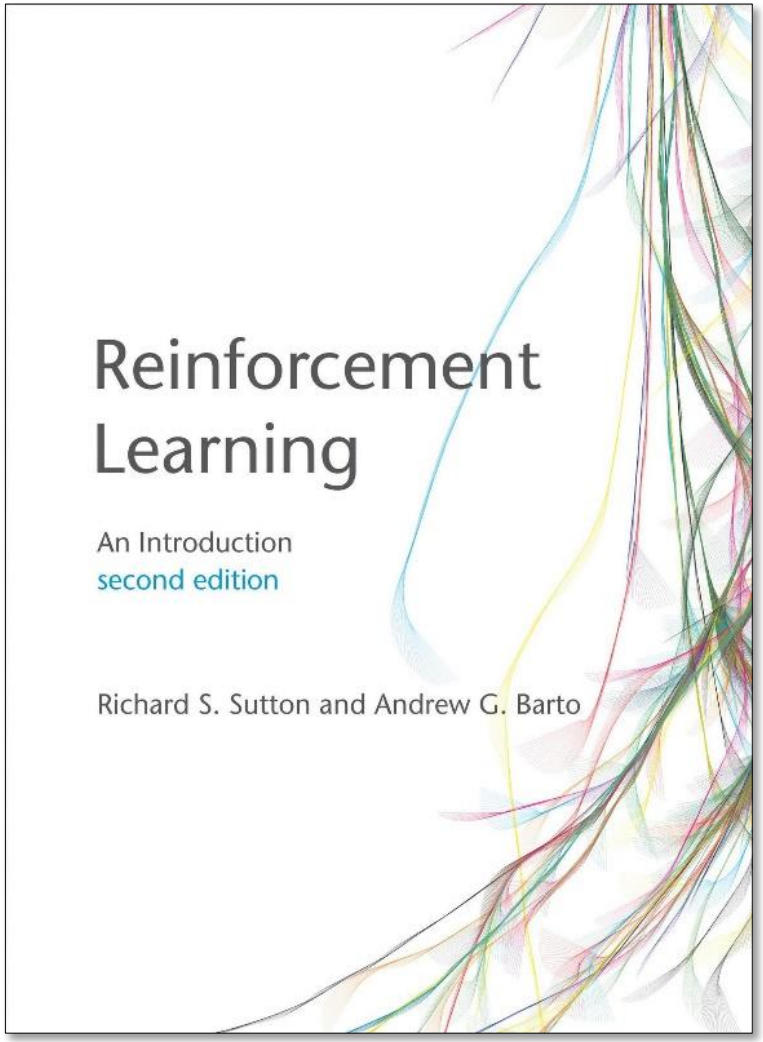
Rich Sutton

March 13, 2019

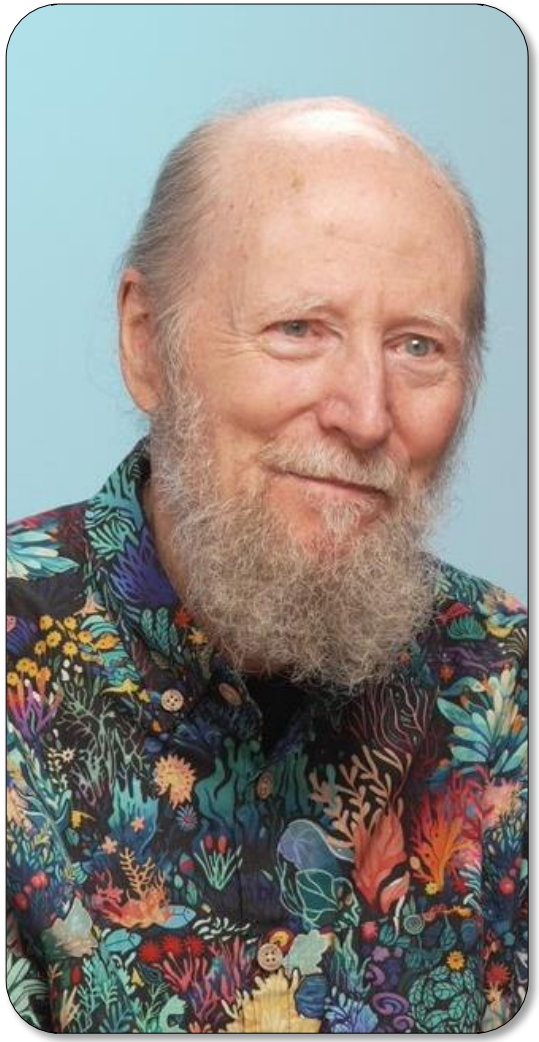
The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin. The ultimate reason for this is Moore's law, or rather its generalization of continued exponentially falling cost per unit of computation. Most AI research has been conducted as if the computation available to the agent were constant (in which case leveraging human knowledge would be one of the only ways to improve performance) but, over a slightly longer time than a typical research project, massively more computation inevitably becomes available. Seeking an improvement that makes a difference in the shorter term, researchers seek to leverage their human knowledge of the domain, but the only thing that matters in the long run is the leveraging of computation. These two need not run counter to each other, but in practice they tend to. Time spent on one is time not spent on the other. There are psychological commitments to investment in one approach or the other. And the human-knowledge approach tends to complicate methods in ways that make them less suited to taking advantage of general methods leveraging computation. There were many examples of AI researchers' belated learning of this bitter lesson, and it is instructive to review some of the most prominent.

In computer chess, the methods that defeated the world champion, Kasparov, in 1997, were based on massive, deep search. At the time, this was looked upon with dismay by the majority of computer-chess researchers who had pursued methods that leveraged human understanding of the special structure of chess. When a simpler, search-based approach with special hardware and software proved vastly more effective, these human-knowledge-based chess researchers were not good losers. They said that ``brute force" search may have won this time, but it was not a general strategy, and anyway it was not how people played chess. These researchers wanted methods based on human input to win and were disappointed when they did not.

incompleteideas.net/IncIdeas/BitterLesson.html



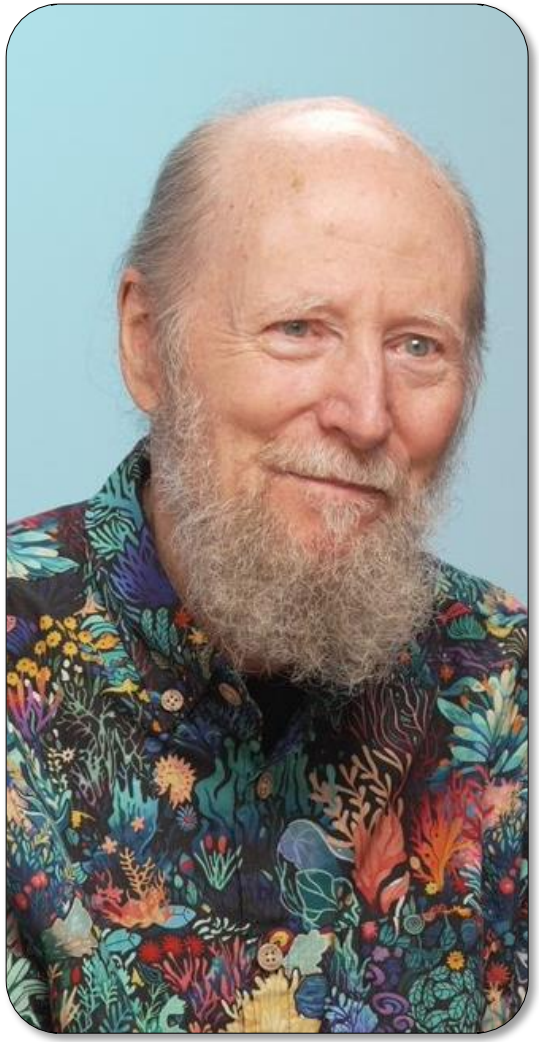
Sutton literally wrote *THE BOOK* on RL



Richard Sutton

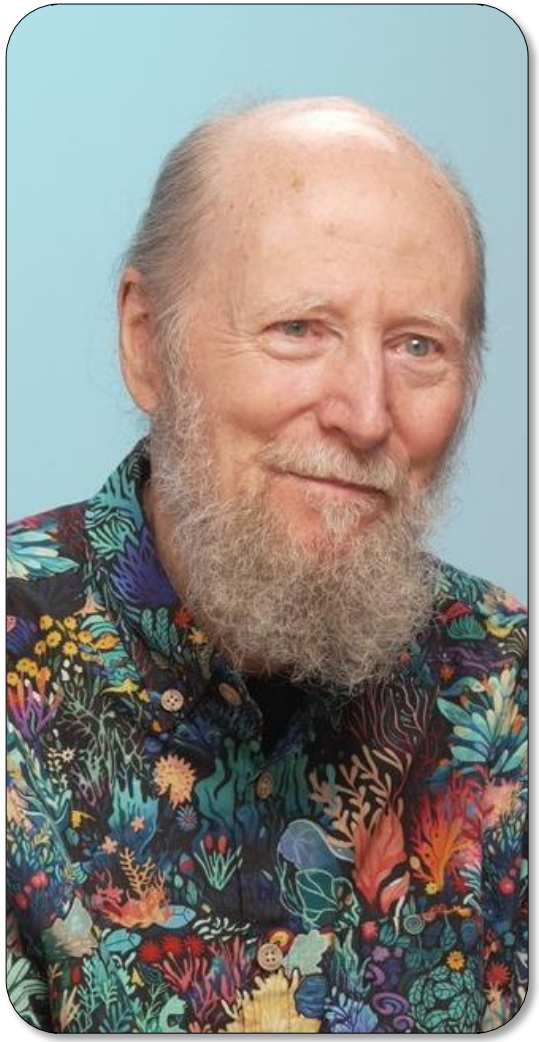
We have to learn the bitter lesson that building in how we think we think does not work in the long run. The bitter lesson is based on the historical observations that 1) AI researchers have often tried to build knowledge into their agents, 2) this always helps in the short term, and is personally satisfying to the researcher, but 3) in the long run it plateaus and even inhibits further progress, and 4) breakthrough progress eventually arrives by an opposing approach based on scaling computation by search and learning. The eventual success is tinged with bitterness, and often incompletely digested, because it is success over a favored, human-centric approach.

incompleteideas.net/IncIdeas/BitterLesson.html



Suttons Kritik an LLMs

Interview mit Dwarkesh Patel:
„Richard Sutton – Father of RL thinks LLMs are a dead end“



Sutton: What is intelligence? The problem is to understand your world. Reinforcement learning is about understanding your world, whereas large language models are about mimicking people, doing what people say you should do. They’re not about figuring out what to do.

Patel: You would think that to emulate the trillions of tokens in the corpus of Internet text, you would have to build a world model. In fact, these models do seem to have very robust world models. They’re the best world models we’ve made to date in AI, right? What do you think is missing?

Sutton: I would disagree with most of the things you just said. To mimic what people say is not really to build a model of the world at all. You’re mimicking things that have a model of the world: people. I would question the idea that [LLMs] have a world model. A world model would enable you to predict what would happen. They have the ability to predict what a person would say. They don’t have the ability to predict what will happen.

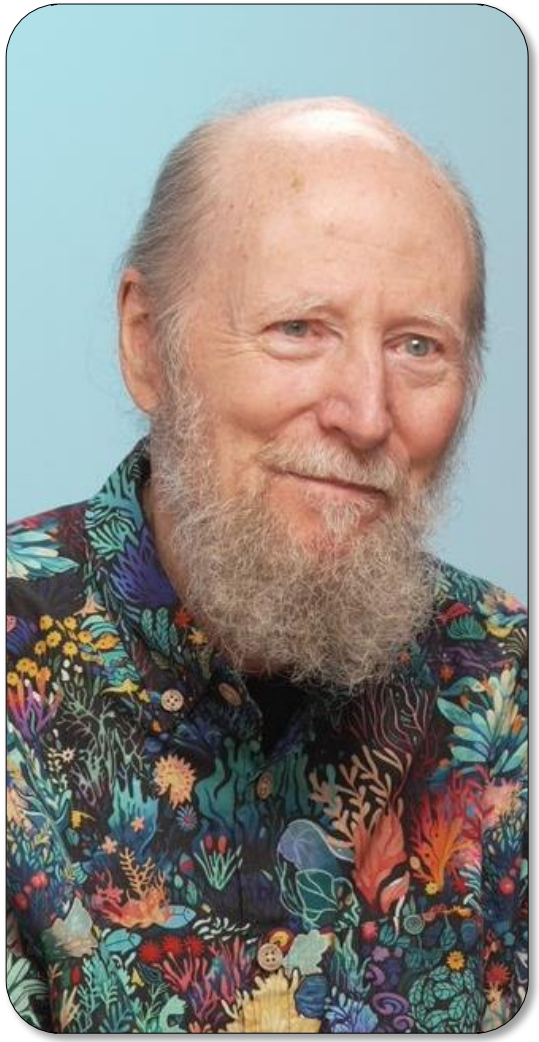
Interview mit Dwarkesh Patel (2025): *Richard Sutton – Father of RL thinks LLMs are a dead end* | https://www.youtube.com/watch?v=21EYKqUsPfg

Suttons Kritik an LLMs

Interview mit Dwarkesh Patel:
„Richard Sutton – Father of RL thinks LLMs are a dead end“

Sutton: What we want, to quote Alan Turing, is a machine that can learn from experience, where experience is the things that actually happen in your life. You do things, you see what happens, and that’s what you learn from. The large language models learn from something else.

I don’t think learning is really about training. I think learning is about learning, it’s about an active process. The child tries things and sees what happens.



Interview mit Dwarkesh Patel (2025): *Richard Sutton – Father of RL thinks LLMs are a dead end* | https://www.youtube.com/watch?v=21EYKqUsPfg

Suttons Kritik an LLMs

Welcome to the Era of Experience

David Silver, Richard S. Sutton*

Abstract

We stand on the threshold of a new era in artificial intelligence that promises to achieve an unprecedented level of ability. A new generation of agents will acquire superhuman capabilities by learning predominantly from experience. This note explores the key characteristics that will define this upcoming era.

Silver & Sutton (2025): *Welcome to the Era of Experience*. Preprint of a chapter to appear in *Designing an Intelligence*, G. Konidaris (Ed.), MIT Press.

Suttons Kritik an LLMs

The Era of Human Data

Artificial intelligence (AI) has made remarkable strides over recent years by training on massive amounts of human-generated data and fine-tuning with expert human examples and preferences. This approach is exemplified by large language models (LLMs) that have achieved a sweeping level of generality. A single LLM can now perform tasks spanning from writing poetry and solving physics problems to diagnosing medical issues and summarising legal documents.

However, while imitating humans is enough to reproduce many human capabilities to a competent level, this approach in isolation has not and likely cannot achieve superhuman intelligence across many important topics and tasks. In key domains such as mathematics, coding, and science, the knowledge extracted from human data is rapidly approaching a limit. The majority of high-quality data sources - those that can actually improve a strong agent's performance - have either already been, or soon will be consumed. The pace of progress driven solely by supervised learning from human data is demonstrably slowing, signalling the need for a new approach. Furthermore, valuable new insights, such as new theorems, technologies or scientific breakthroughs, lie beyond the current boundaries of human understanding and cannot be captured by existing human data.

Silver & Sutton (2025): *Welcome to the Era of Experience*. Preprint of a chapter to appear in *Designing an Intelligence*, G. Konidaris (Ed.), MIT Press.

Suttons Kritik an LLMs

The Era of Experience

To progress significantly further, a new source of data is required. This data must be generated in a way that continually improves as the agent becomes stronger; any static procedure for synthetically generating data will quickly become outstripped. This can be achieved by allowing agents to learn continually from their own *experience*, i.e., data that is generated by the agent interacting with its environment. AI is at the cusp of a new period in which experience will become the dominant medium of improvement and ultimately dwarf the scale of human data used in today's systems.

Silver & Sutton (2025): *Welcome to the Era of Experience*. Preprint of a chapter to appear in *Designing an Intelligence*, G. Konidaris (Ed.), MIT Press.

LeCuns Kritik an (autoregressiven) LLMs

ESSAY TECHNOLOGY & THE HUMAN

BY JACOB BROWNING AND YANN LECUN

AUGUST 23, 2022

AI And The Limits Of Language

An artificial intelligence system trained on words and sentences alone will never approximate human understanding.

https://www.noemamag.com/ai-and-the-limits-of-language/

LeCuns Kritik an autoregressiven LLMs

Yann LeCun (2025): *Self-Supervised Learning, JEPA, World Models, and the Future of AI*

We’re nowhere near matching human intelligence or even animal intelligence with the type of techniques that we have access to at the moment.

[…]

We are never going to get to human level AI just by training on text. It’s not going to happen despite what you might hear from some of the more optimistic sounding CEOs of various AI companies in Silicon Valley.

Vortrag am Center of Mathematical Sciences and Applications, Harvard University | https://www.youtube.com/watch?v=yUmDRxVokrg

LeCuns Kritik an autoregressiven LLMs

Yann LeCun (2025): *Self-Supervised Learning, JEPA, World Models, and the Future of AI*

We have systems that can pass the bar exam, they can solve math problems, you know, do all kinds of stuff that is intellectually challenging for most of us. But we still don't have robots that can do what a cat can do, or what a 10-year-old can do [on its first try]. You tell a 10-year-old, clear out the dinner table and fill the dishwasher. A 10-year-old can do it without being trained to do it. Basically the first time.

[...]

So, obviously we're missing something big.

Vortrag am Center of Mathematical Sciences and Applications, Harvard University | https://www.youtube.com/watch?v=yUmDRxVokrg

LeCuns Kritik an autoregressiven LLMs

Yann LeCun (2025): *Self-Supervised Learning, JEPA, World Models, and the Future of AI*

The other issue also with current architectures is that they use autoregressive prediction. So they use their own predictions as input to make further predictions, and that leads to divergence or hallucination as people call it. So there's a lot of things that really we are missing to kind of match the type of intelligence we observe in humans and animals. Humans and animals have ***mental models of the world***. The behavior is driven by objectives, by tasks, goals, if you want, they can reason and they can plan complex action sequences. All things that chat bots and LLMs are essentially incapable of, or at least not to the level that we'd like.

Vortrag am Center of Mathematical Sciences and Applications, Harvard University | https://www.youtube.com/watch?v=yUmDRxVokrg

LeCuns Kritik an autoregressiven LLMs

Yann LeCun (2025): *Self-Supervised Learning, JEPA, World Models, and the Future of AI*

So the main issue is that current AI architectures and machine learning techniques suck compared to what we can observe in humans and animals, the type of efficiency in learning that we see in animals and humans.

[...]

Early on in machine learning, the main technique was supervised learning, and then there was a big fashion around reinforcement learning for a while. Now it's used a lot of course to fine tune large language models, but in themselves, those two techniques are really insufficient. The type of learning that we observe in humans and animals is very different. It's neither supervised nor reinforced for that matter. It's more like ***self-supervised learning***.

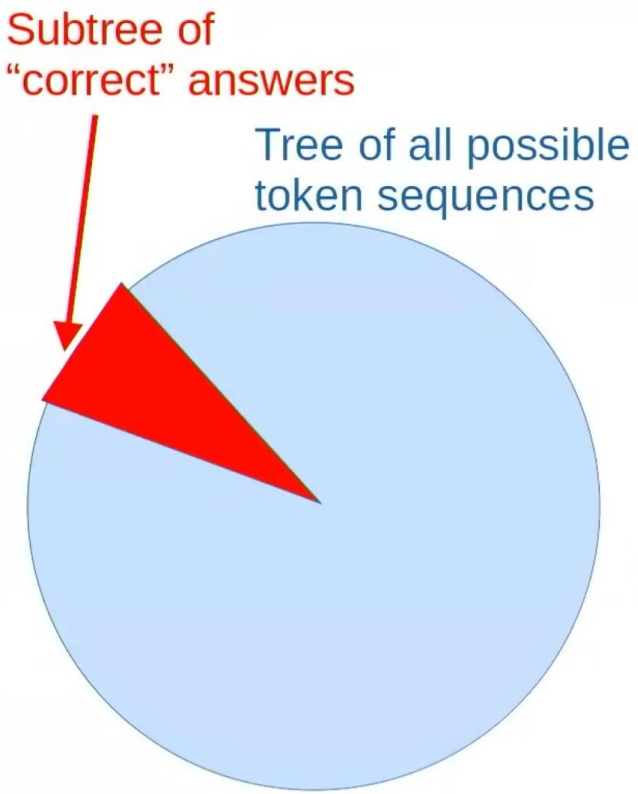
Vortrag am Center of Mathematical Sciences and Applications, Harvard University | https://www.youtube.com/watch?v=yUmDRxVokrg

LeCuns Kritik an autoregressiven LLMs

Yann LeCun (2025): *Self-Supervised Learning, JEPA, World Models, and the Future of AI*

Auto-Regressive Generative Models Suck!

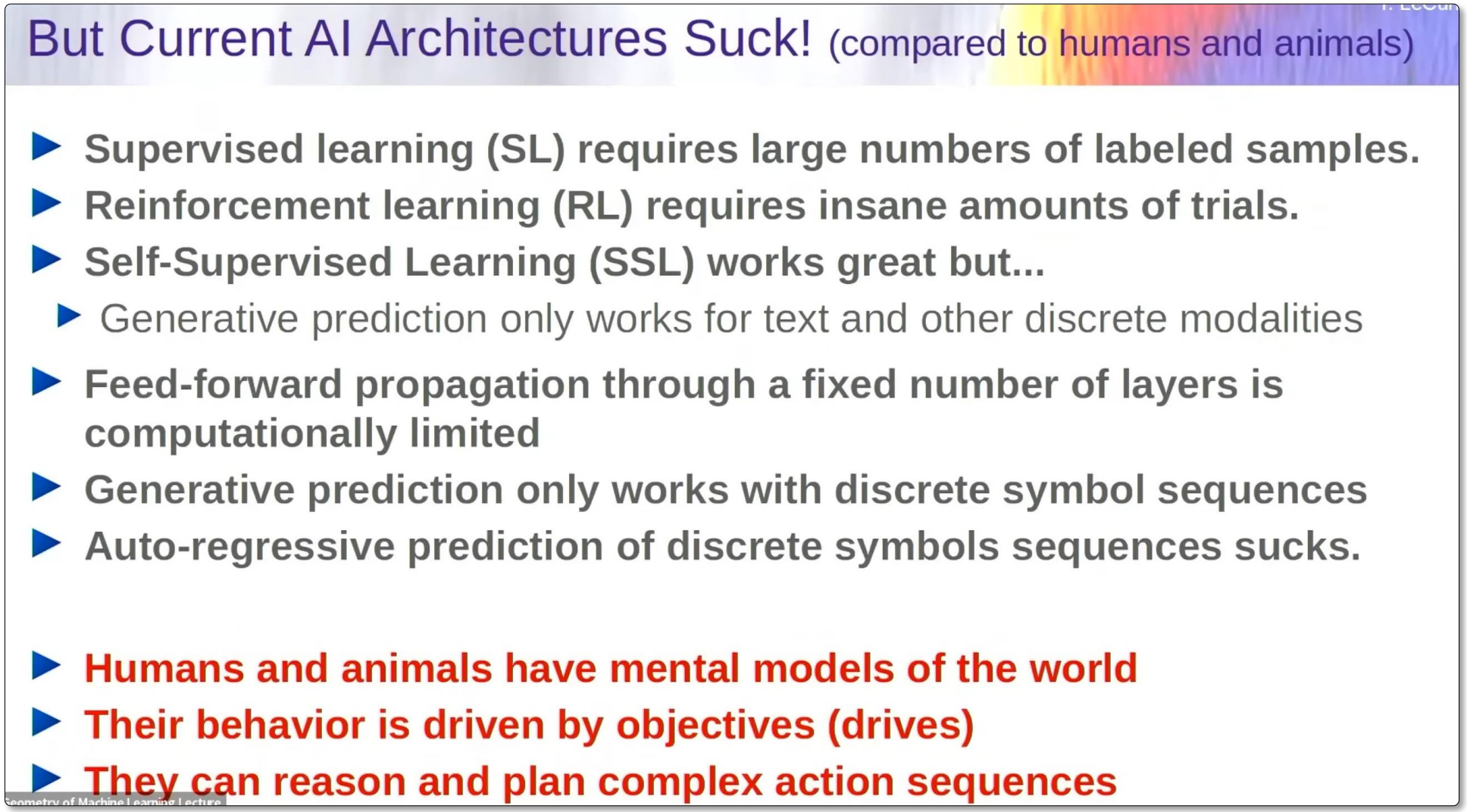
- Auto-Regressive LLMs are **doomed**.
- They cannot be made factual, non-toxic, etc.
- They are not controllable
- Probability e that any produced token takes us outside of the set of correct answers
- Probability that answer of length n is correct (assuming independence of errors):
  - $P(\text{correct}) = (1-e)^n$
- **This diverges exponentially.**
- **It's not fixable (without a major redesign).**
- See also [Dziri...Choi, ArXiv:2305.18654]



Vortrag am Center of Mathematical Sciences and Applications, Harvard University | https://www.youtube.com/watch?v=yUmDRxVokrg

LeCuns Kritik an autoregressiven LLMs

Yann LeCun (2025): *Self-Supervised Learning, JEPA, World Models, and the Future of AI*



Vortrag am Center of Mathematical Sciences and Applications, Harvard University | https://www.youtube.com/watch?v=yUmDRxVokrg

LeCuns Kritik an autoregressiven LLMs

Yann LeCun (2025): *Self-Supervised Learning, JEPA, World Models, and the Future of AI*

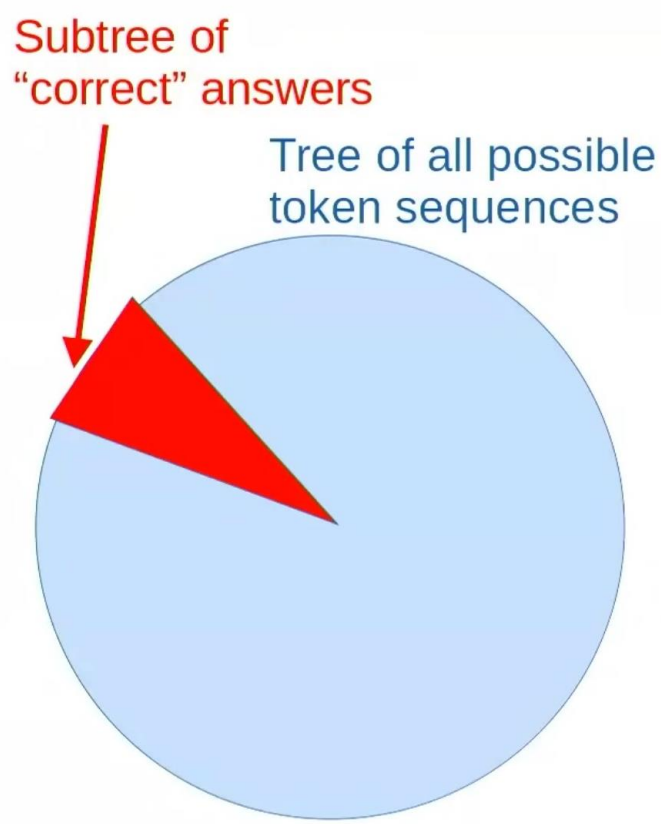
- AI systems that can pass the bar exam, do math problems, prove theorems….
- …but where are my Level-5 self-driving car and my domestic robot?
- We keep bumping into Moravec's paradox
  - Things that are easy for humans are difficult for AI and vice versa.

Vortrag am Center of Mathematical Sciences and Applications, Harvard University | https://www.youtube.com/watch?v=yUmDRxVokrg

LeCuns Kritik an autoregressiven LLMs

Autoregressive LLMs erzeugen Text Token für Token. Jeder neue Schritt hängt vom bisherigen Output ab.

- Problem: Ein kleiner Fehler an einer Stelle kann den weiteren Verlauf „beschädigen“.
- Alle möglichen Token-Folgen bilden einen riesigen Baum.
- Die korrekten Antworten sind nur ein kleiner Teilbaum.
- Wenn das Modell an einer Stelle falsch abbiegt, landet es außerhalb des korrekten Bereichs.
- LeCuns starke These: Bei längeren Antworten wächst das Fehlerrisiko systematisch.
- Grundproblem: Fehler können sich über viele Schritte exponentiell aufschaukeln.



„World Modells“

Intuitiv: Ein *World-Modell* ist eine gelernte innere „Simulation“ der Welt: Das System lernt, welche Dinge typischerweise passieren, was möglich, wahrscheinlich oder unmöglich ist, und was aus bestimmten Handlungen folgen könnte. Zweck: vorauszuschauen, planen, Fehler vermeiden, neue Situationen mit weniger *Trial-and-Error* bewältigen.

Im Reinforcement Learning: Ein World-Modell ist ein Modell der Umgebung, das dem *Agent* sagt oder schätzt, was nach einer Handlung in einem Zustand passiert: also typischerweise den nächsten Zustand und die Belohnung, oft als Wahrscheinlichkeitsverteilung. Es dient dazu, Handlungsfolgen intern zu simulieren, *Policies* zu bewerten und zu planen, statt alles durch reale Interaktion ausprobieren zu müssen.

Bei LeCun / JEPA: Ein World-Modell ist ein selbstüberwacht gelerntes, prädiktives Modell, das nicht primär Pixel oder Wörter rekonstruiert, sondern abstrakte Repräsentationen der Welt vorhersagt. In JEPA/V-JEPA werden verdeckte oder zukünftige Teile einer Videoszene im latenten Raum vorhergesagt; dadurch soll das System relevante Struktur, Physik, Objektinteraktionen und zeitliche Entwicklung lernen. Ziel ist ein Modell, das Common Sense, effizientes Lernen, hierarchische Planung und Handeln unter Unsicherheit ermöglicht.

Vision Language Action Models (VLAs)

VLAs sind multimodale Modelle für Robotersteuerung, die oft LLMs (oder zumindest LLM-artige Systeme) als Komponenten beinhalten.

Manche nutzen explizit vortrainierte *LLMs* oder *Vision-Language*-Modelle und erweitern sie so, dass sie nicht nur Text, sondern auch Roboteraktionen ausgeben können.

Als Kontrastfolie ist es hilfreich, zunächst einen kurzen Blick darauf zu werfen, was Prä-LLM-Ära Roboter (beispielsweise in der Industrie) gut können und wo ihre Grenzen liegen.

Klassische Industrieroboter

Typischer Einsatz: Fabriken, Montage, Schweißen, Lackieren, Verpackung

Aufgabe meist eng definiert: gleicher Ort, gleiche Objekte, gleicher Ablauf, kontrollierte Umgebung.

Programmierung: Bewegungspfade, Regeln, Sensor-Signale, Wenn-dann-Abläufe.

Stärke: schnell, präzise, zuverlässig bei wiederholbaren Aufgaben.

Aber: kaum flexibel, nicht ohne aufwendige Neuprogrammierung für neue Aufgaben einsetzbar.

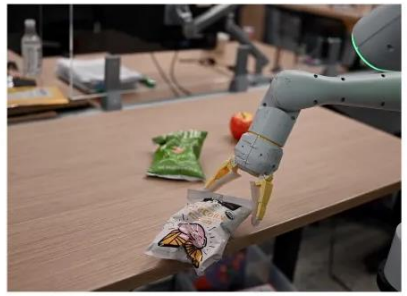
Insbesondere: Kann typischerweise keine Anweisungen in einem breiten Spektrum natürlicher Sprache entgegennehmen.

RT-2: Move the coke can to Taylor Swift

https://deepmind.google/blog/rt-2-new-model-translates-vision-and-language-into-action/

RT-2: Move the coke can to Taylor Swift

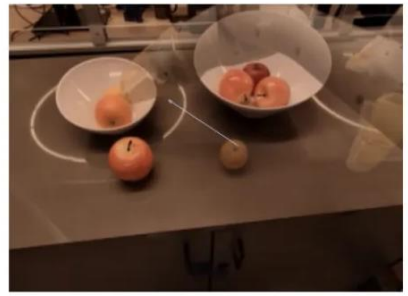
put strawberry into the correct bowl



pick up the bag about to fall off the table

move apple to Denver Nuggets

pick robot



place orange in matching bowl

move Red Bull can to H

move soccer ball to basketball

move banana to Germany

move cup to the wine bottle

pick animal with different colour

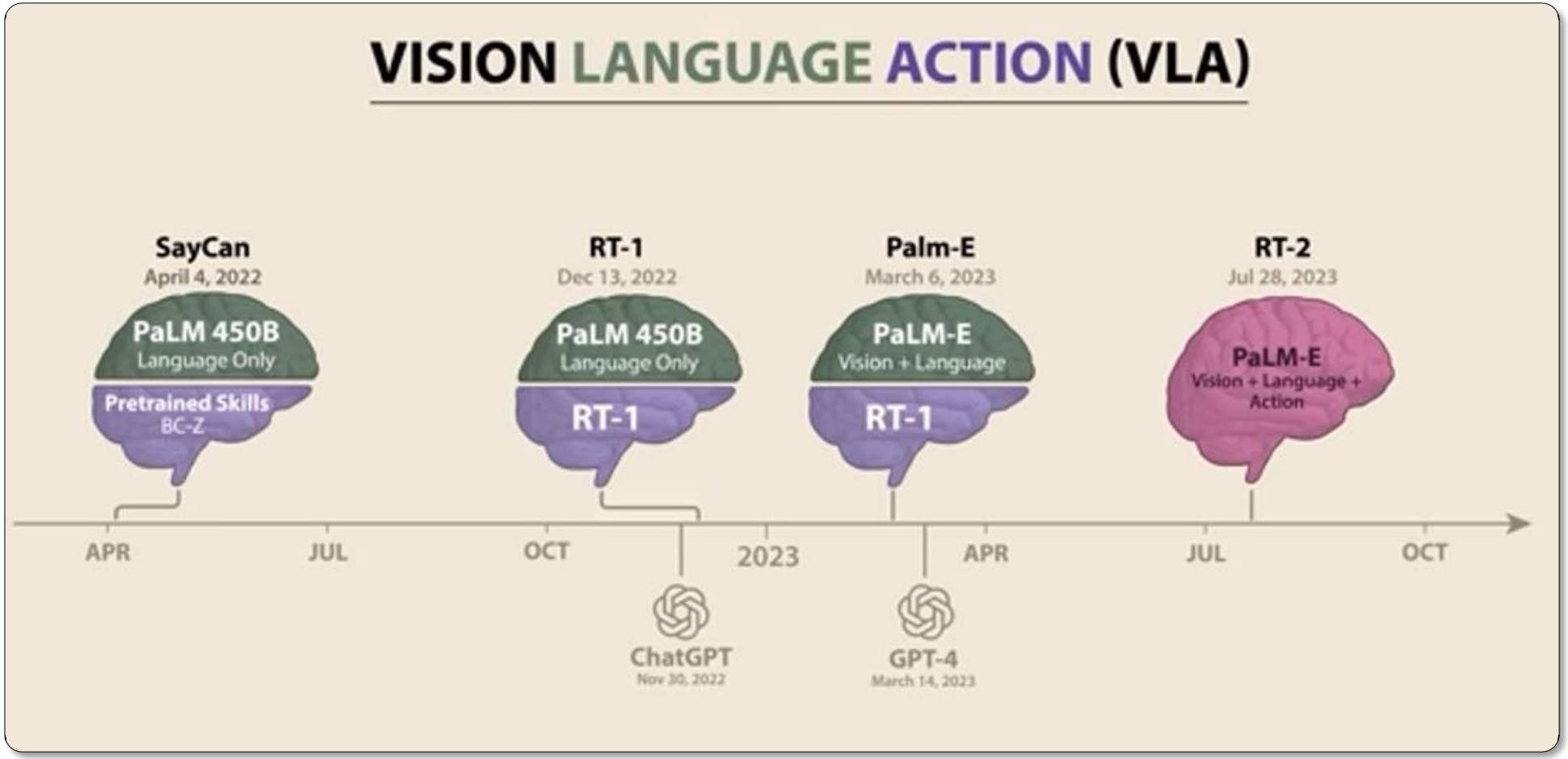
https://deepmind.google/blog/rt-2-new-model-translates-vision-and-language-into-action/

Vision Language Action Models (VLAs)

Vision Language Action Models (VLAs)

Welch Labs (2026): *Inside the World's Smartest Robot Brain [VLA]* | https://www.youtube.com/watch?v=2mrGMMmrVNE

Vision Language Action Models (VLAs)



Welch Labs (2026): *Inside the World's Smartest Robot Brain [VLA]* | https://www.youtube.com/watch?v=2mrGMMmrVNE