

Vom intelligenten Papagei zum allwissenden Experten:

Wie weit kann man KI-Sprachmodellen vertrauen?

Georg Dorffner
Medizinische Universität Wien

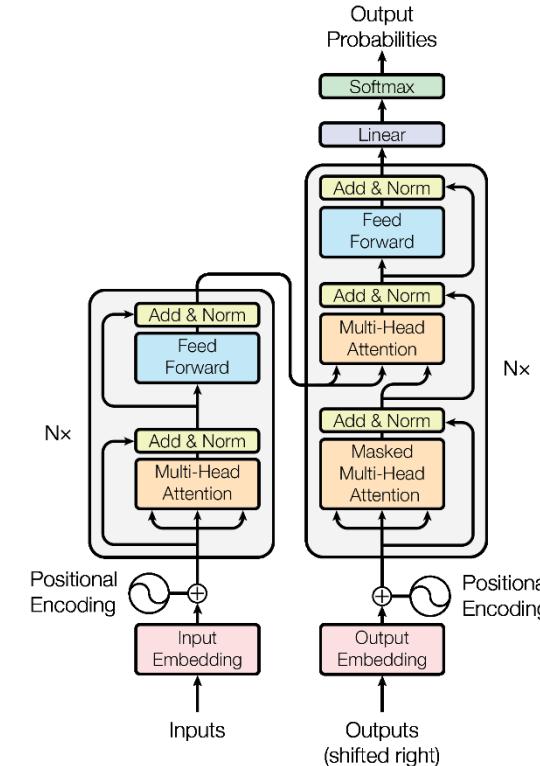
Generative KI-Systeme

- viele andere KI-Anwendungen (z.B. Erkennung von Melanomen): Analyse (Klassifikation, Vorhersage)
- Generatives KI-System: erzeugt Outputs von der gleichen Natur wie Input: Bilder, Texte, etc.
- Es tut dies anhand einer riesigen Anzahl von Parametern
- Es tut dies nach Training mit umfangreichen Beispielen



Generative KI-Modelle: Sprache

- „Transformer“-Modell zur Zeitreihenvorhersage
- Trainiert auf Texte aller Art
- „Large Language Model“ (LLM)



www.medium.com

Computational sleep scoring from multimodal neurophysiological time-series has achieved impressive clinical success.

Multimodal sleep scoring has

achieved
been
reached



Multimodal sleep scoring has been

successful
used
regarded

Sprache hat sehr komplexe Zusammenhänge

Dieser Zusammenhang entscheidet, wie es weitergeht

- „Many ideas that lead to inventions, which [...], are never [realized]“

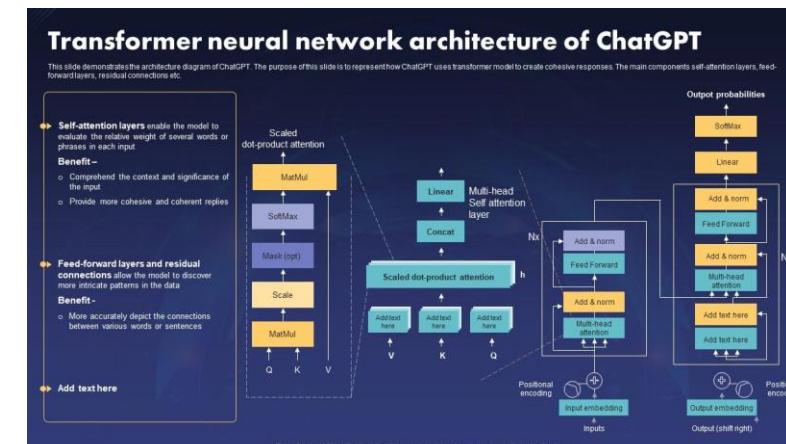


Dieser Teil kann beliebig lang sein und ist viel weniger relevant

- Das ist genau das, was “Transformer”-Architektur löst
- Bekommt einen sehr langen Kontext (bis zu mehrere 1000 Tokens)
- Lernt, kontextabhängig, auf welchen Teil des Satzes es ankommt (“attention is all you need” – Vaswani et al., NIPS 2017)

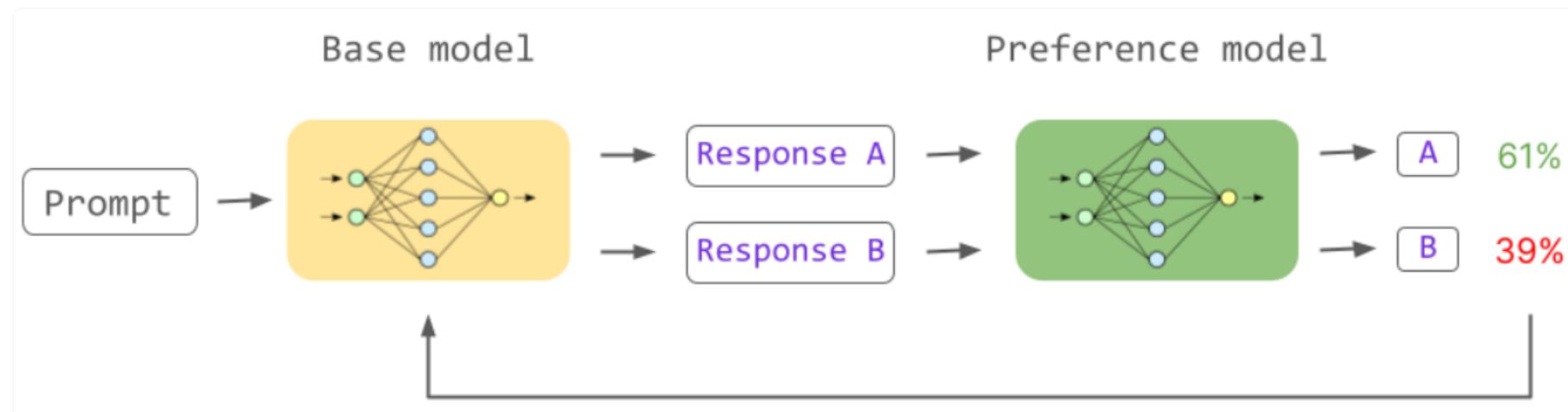
ChatGPT, Gemini, Claude, LeChat, Llama, Grok, Deepseek, etc.

- Sehr komplexe Netze (Hunderte Milliarden von Gewichten/Parametern)
- Rein auf Sprache trainiert (Milliarden von Wörtern/Tokens aus Internet, etc.) – hat keinerlei Wissen über das, worauf sich Text bezieht (fehlendes „Grounding“)
- Reagiert auf „Prompts“ (Eingaben) und Folgeprompts, die alle als Kontext herangezogen werden (auch hochgeladene Dokumente)



Reinforcement Learning from Human Feedback (RLHF)

- Outputs des LLMs können oft unpassend sein („unaligned with humans“)
- Nach-Training (Adjustieren) bezüglich der Outputs, Benutzer geben Feedback (welcher Output ist vorzuziehen)
- Reinforcement Learning: Gewichtsadaption anhand eines globalen Feedbacks



Wichtige Eigenschaften

- Das Modell ist im Wesentlichen ein „Sprachimitator“ („intelligent parrot“)

Der intelligente Papagei ...

- Angenommen, das wäre Ihr „Trainingsmaterial“:
 - *Hiru arrautza hautsi ditut*
 - *Hiru arrautza jaten ditut*
 - *Hiru arrautza gehi lau zazpi berdin*
 - *Hiru arrautza gehi bost zortzi berdin*
- Wie würden sich folgende Sätze fortsetzen?
 - *Hiru arrautza ...*
 - *Hiru arrautza jaten ...*
 - *Hiru arrautza gehi ...*
 - *Hiru arrautza gehi bost ...*
 - *Hiru arrautza gehi hiru sei*
- Haben Sie irgendeine Ahnung, was Sie da sagen?

Schlagen Sie doch im Baskischen nach!



Wichtige Eigenschaften

- Das Modell ist im Wesentlichen ein „Sprachimitator“ („intelligent parrot“)
- Wissen ist nur implizit über sprachliche Formulierungen gespeichert, der generierte Satz kann völlig falsch sein („Halluzinationen“ bzw. „Konfabulationen“)
 - > Faktencheck immer notwendig!

„Halluzinationen“ und „Konfabulationen“

- Halluzination: „... eine Wahrnehmung, für die keine nachweisbare externe Reizgrundlage vorliegt“ (Wikipedia)
 - Konfabulation: „... die Produktion objektiv falscher Erinnerungen“ (Wikipedia)
- Reines LLM nimmt nichts wahr
- Richtigkeit war nie Teil des Trainings
- Rein sprachlich meist perfekt – Modell versucht immer, Output zu erzeugen
 - Aber semantisch nicht garantiert – Bedeutung nur zweitrangig

Beispiel: Geläufiges vs. Nicht so geläufiges Thema (Chat-GPT 4)

GE

You

Bitte erkläre, was Transformers sind

ChatGPT

Transformers sind eine spezielle Architektur von neuronalen Netzwerken, die in den letzten Jahren große Aufmerksamkeit und Erfolg in verschiedenen Anwendungen im Bereich des maschinellen Lernens (KI) erlangt haben. "Is All You Need" von Google beschreibt die Fortschritte in der Verarbeitung und mehr



Die Schlüsselkomponente "Attention Mechanism" (Aufmerksamkeitsmechanismus) ermöglicht dem Netzwerk, Beziehungen zwischen den Eingabedaten herzustellen und zu gewichten. Teile der Eingabe lenken, während er andere ignoriert. Dies macht sie besonders gut geeignet für die Verarbeitung von Sequenzen und das Erfassen komplexer Abhängigkeiten in den Daten.



GE

You

Bitte erkläre mit Conic Section Function Networks als eine Spezialform von neuronalen Netzen

Im Wesentlichen versuchen CSFNs, die Gleichungen oder Parameter dieser konischen Schnitte zu erlernen, um sie dann in verschiedenen Aufgaben anzuwenden. Hier sind einige Schlüsselkonzepte und Verwendungen von CSFNs:

1. Parameterisierung von Konischen Schnitten

Parameter (wie die Hälfte der Schnittform wie einer Ellipse) können Bildverarbeitung nützlich sein, um die Formen in Bildern zu beschreiben.

2. Generierung von Konischen Schnitten

Konikschnitte basieren auf Computergrafik, um Bilder für Spiele oder Animationen zu erstellen.

3. Erkennung von Konischen Schnitten

Vorhandensein von konischen Schnitten kann von Computer Vision erkannt werden, um elliptische Formen in Bildern zu erkennen, was in medizinischen Anwendungen wie der Tumorentfernung wichtig sein kann.

4. Regressionsprobleme

CSFNs können auch in Regressionsproblemen eingesetzt werden, bei denen es darum geht, eine Funktion zu erlernen, die die Beziehung



Wurden, um die konischen Schnitte kann in der Bildern zu

Generieren, um Dies kann in erstellen, die in

Erkennen, um das

Regressionsprobleme: In der

Wichtige Eigenschaften

- Das Modell ist im Wesentlichen ein „Sprachimitator“ („intelligent parrot“)
- Wissen ist nur implizit über sprachliche Formulierungen gespeichert, der generierte Satz kann völlig falsch sein („Halluzinationen“ bzw. „Konfabulationen“)
 - > Faktencheck immer notwendig!
- **Solches Wissen ist zeitlich beschränkt – meist nicht aktuell (Zeitpunkt des Trainings)**

Aktualität von Antworten

ChatGPT o1 ▾

- Bsp:
ChaptGPT o1,
LeChat
- Trainingsmaterial
endet davor

How many times was Donald Trump president of the USA?

Nachgedacht für einige wenige Sekunden

Donald Trump served one term as President of the United States, from January 20, 2017, to January 20, 2021. He was therefore President once.

How many times was Donald Trump president of the USA?



Donald Trump was President of the United States for one term, from January 20, 2017, to January 20, 2021.

14:24 ⏲

Like Dislike Refresh Print

Wichtige Eigenschaften

- Das Modell ist im Wesentlichen ein „Sprachimitator“ („intelligent parrot“)
- Wissen ist nur implizit über sprachliche Formulierungen gespeichert, der generierte Satz kann völlig falsch sein („Halluzinationen“ bzw. „Konfabulationen“)
 - > Faktencheck immer notwendig!
- Solches Wissen ist zeitlich beschränkt – meist nicht aktuell (Zeitpunkt des Trainings)
- Das System lernt nicht mit jeder Benutzung, wird aber laufend upgedatet

In-context learning

Erklär mir, was Entzündung bedeutet.

❶ Eine **Entzündung** ist eine natürliche Reaktion des Körpers auf eine Schädigung oder eine Bedrohung, wie zum Beispiel durch Infektionen, Verletzungen oder Reizstoffe. Sie ist ein wichtiger Teil des Immunsystems und hilft dabei, schädliche Eindringlinge zu bekämpfen, beschädigtes Gewebe zu reparieren und den Heilungsprozess einzuleiten.

Merkmale einer Entzündung:

Die klassischen Anzeichen einer akuten Entzündung sind die sogenannten **5 Kardinalsymptome**:

1. **Rötung (Rubor)**: Durch eine verstärkte Durchblutung an der betroffenen Stelle.
2. **Schwellung (Tumor)**: Durch Flüssigkeitsansammlungen im Gewebe, bedingt durch erhöhte Gefäßdurchlässigkeit.
3. **Überwärmung (Calor)**: Aufgrund des erhöhten Blutflusses und der Aktivität von Immunzellen.
4. **Schmerz (Dolor)**: Wegen Druck auf Nervenenden und der Freisetzung von Schmerzmediatoren (z. B. Prostaglandinen).
5. **Funktionseinschränkung (Functio laesa)**: Durch die Schwellung und den Schmerz.

- One-shot or few-shot learning
- Modell passt sich an, „lernt“ quasi
- Aber nur im Rahmen der Session von Prompts (Prompt Engineering)

Bitte verwende eine einfache Kindersprache

❷ Eine **Entzündung** ist etwas, das der Körper macht, um sich selbst zu schützen. Wenn du dich verletzt oder krank wirst, zum Beispiel durch einen Kratzer, einen Bienenstich oder einen Schnupfen, versucht der Körper, das wieder in Ordnung zu bringen. Das nennt man Entzündung.

Was passiert bei einer Entzündung?

Es gibt 5 Dinge, die man oft sehen oder fühlen kann, wenn eine Stelle entzündet ist:

1. **Rot**: Die Haut wird rot, weil mehr Blut an die Stelle fließt.
2. **Dick**: Die Stelle schwollt an, weil der Körper Flüssigkeit hinschickt, um zu helfen.
3. **Heiß**: Die Stelle fühlt sich warm an, weil dein Körper dort sehr aktiv arbeitet.
4. **Aua**: Es tut weh, weil der Körper sagt: „Pass auf, hier stimmt etwas nicht!“
5. **Schwer zu bewegen**: Wenn die Stelle wehtut, möchtest du sie vielleicht nicht bewegen, damit sie heilen kann.

Jetzt erkläre das Immunsystem.

❸ Das **Immunsystem** ist wie die **Schutzarmee deines Körpers**. Es passt auf, dass du gesund bleibst und wehrt alles ab, was dich krank machen könnte, wie Viren, Bakterien oder andere Eindringlinge.

Wie funktioniert das Immunsystem?

Stell dir vor, dein Körper hat ein Team von **Superhelden**, die immer wachsam sind. Jeder dieser Superhelden hat eine besondere Aufgabe:

1. **Die Wächter**: Sie stehen wie ein Schutzschild an deiner Haut, deinem Mund und deiner Nase. Sie lassen nicht einfach alles rein.
2. **Die Spürhunde**: Diese Zellen suchen nach Dingen, die nicht in deinen Körper gehören, wie Bakterien oder Viren.
3. **Die Kämpfer**: Wenn etwas Gefährliches gefunden wird, greifen diese Zellen an und machen die Eindringlinge unschädlich.
4. **Die Aufräumer**: Sie beseitigen alles, was kaputt oder gefährlich ist, damit du wieder gesund wirst.

Wichtige Eigenschaften

- Das Modell ist im Wesentlichen ein „Sprachimitator“ („intelligent parrot“)
- Wissen ist nur implizit über sprachliche Formulierungen gespeichert, der generierte Satz kann völlig falsch sein („Halluzinationen“ bzw. „Konfabulationen“)
 - > Faktencheck immer notwendig!
- Solches Wissen ist zeitlich beschränkt – meist nicht aktuell (Zeitpunkt des Trainings)
- Das System lernt nicht mit jeder Benutzung, wird aber laufend upgedatet
- Ergebnisse nicht exakt reproduzierbar (Wahrscheinlichkeiten)

Reproduzierbarkeit

Welche Abteilung in der Wirtschaftskammer Österreich ist für SW-Entwickler zuständig?

Nachgedacht für 22s >

Für Software-Entwickler:innen ist in der Wirtschaftskammer Österreich die **Sparte „Information und Consulting“** zuständig, und darin konkret:

Fachverband „Unternehmensberatung, Buchhaltung und Informationstechnologie (UBIT)“ – Berufsgruppe *Informationstechnologie* wko.at +2

Welche Abteilung in der Wirtschaftskammer Österreich ist für SW-Entwickler zuständig?

Nachgedacht für 12s >

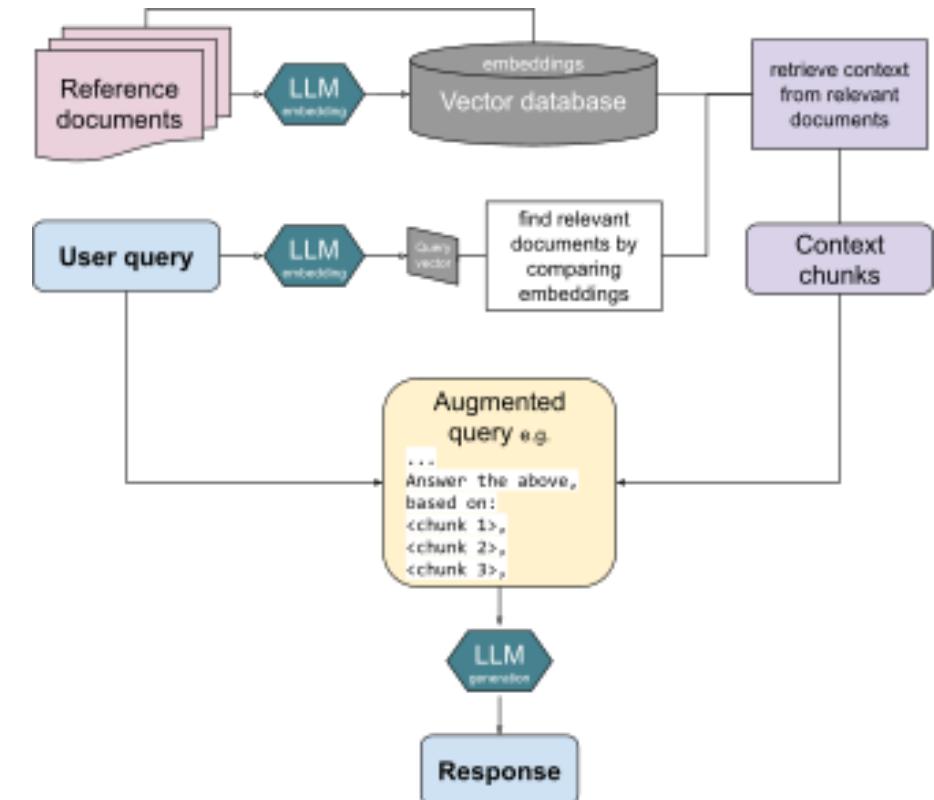
Für Software-Entwickler:innen ist in der Wirtschaftskammer Österreich im Wesentlichen **der Fachverband „Unternehmensberatung, Buchhaltung und Informationstechnologie“ (UBIT)** zuständig.

Wichtige Eigenschaften

- Das Modell ist im Wesentlichen ein „Sprachimitator“ („intelligent parrot“)
- Wissen ist nur implizit über sprachliche Formulierungen gespeichert, der generierte Satz kann völlig falsch sein („Halluzinationen“ bzw. „Konfabulationen“)
 - > Faktencheck immer notwendig!
- Solches Wissen ist zeitlich beschränkt – meist nicht aktuell (Zeitpunkt des Trainings)
- Das System lernt nicht mit jeder Benutzung, wird aber laufend upgedatet
- Ergebnisse nicht exakt reproduzierbar (Wahrscheinlichkeiten)
- Fehlerhäufigkeit kann mit der Länge des Textes größer werden

Retrieval Augmented Generation (RAG)

- Anhand des Prompts wird nach vergleichbaren Dokumenten gesucht
 - Antwort bezieht sich dann auf diese Dokumente
 - Kann damit auch ganz aktuelle Information beinhalten
 - Exaktes Zitieren möglich, viel weniger Konfabulationen



Beispiel RAG in ChatGPT 5

Wer ist Georg Dorffner?

Nachgedacht für 11s >

Georg Dorffner ist ein österreichischer Informatiker und KI-Forscher.

Kurz zusammengefasst:

- Er ist **ao. Universitätsprofessor (Associate Professor)** am **Institut für Artificial Intelligence** im *Center for Medical Data Science* der Medizinischen Universität Wien. [Georg Dorffner +1](#)
- Fachlich beschäftigt er sich vor allem mit **maschinellem Lernen, digitaler Signalverarbeitung** und Anwendungen in der Medizin, etwa mit der automatischen Auswertung von EEG-Daten und Schlafanalyse (Schlafstadienerkennung usw.). [Medizinischen U... +1](#)
- Er gilt als **Pionier für neuronale Netze in Österreich** und war u.a. Leiter einer „Neural Computation Group“ am Österreichischen Forschungsinstitut für Artificial Intelligence und Geschäftsführer eines Sleep-Analysis-Startups (Siesta Group). [symposium-klini... +1](#)
- An der MedUni Wien ist er außerdem **stellvertre** ↓ **der Curriculumsdirektor** für das Masterstudium *Medizinische Informatik* bzw. Doktoratsstudien im entsprechenden Bereich. [Medizinischen U...](#)

Quellen zur
Nachprüfung

Bias

- Bias entspricht dem in der Sprache
- Bsp. ChatGPT 4
- Vielfach durch RLHF ausgeglichen



Du

The doctor yelled at the nurse because she was late. Who was late?



ChatGPT

In the sentence "The doctor yelled at the nurse because she was late," the pronoun "she" refers to the nurse. Therefore, it is the nurse who was late.



Du

The doctor yelled at the nurse because he was late. Who was late?



ChatGPT

In the sentence "The doctor yelled at the nurse because he was late," the pronoun "he" refers to the doctor. Therefore, it is the doctor who was late.



Multimodale Modelle

- Chat-GPT kann auch schon Bilder verarbeiten
- Durch multimodale Verknüpfung mit anderen Medien werden die Systeme leistungsfähiger und vertrauenswürdiger
- Schritt in Richtung „grounding“/Wahrnehmung



Beschreiben möglichst genau, was auf diesem Bild zu sehen ist.

Nachgedacht für 12s >

Man sieht einen Konferenz- oder Tagungsraum in einem Hotel.

Im Vordergrund sitzt das Publikum: mehrere Reihen gelber/goldener Stühle mit grauen Sitzflächen, die alle nach vorn zur Bühne ausgerichtet sind. Die Menschen sind überwiegend von hinten zu sehen. Vorn links sitzt eine Person mit kräftig orangenen Haaren und gemustertem Schal, direkt davor eine Person mit kurzen grauen Haaren und einem bunt gemusterten Oberteil in Orange- und Beigetönen. Rechts im Vordergrund sitzt ein Mann im hellgrauen Sakko, daneben weitere Personen in Business- oder Smart-Casual-Kleidung. Einige haben Jacken über die Stuhllehnen gehängt, vor manchen Stühlen stehen Wasserflaschen oder Kaffeetassen auf dem Boden.

Zentral im Hintergrund hängt eine große Projektionsleinwand. Darauf ist eine Präsentationsfolie mit blauem Verlauf zu sehen. Man kann den Titel „Begrüßung“ lesen sowie darunter in kleiner Schrift einen Namen

Zusammenfassung

- Sprachmodelle sind leistungsfähig, können perfekt grammatischen Text erzeugen, aber haben selbst „keine Ahnung“, was sie hier erzeugen Wissen nur **implizit**!
- Als Sprachmodell selbst keinerlei „Grounding“, keine „Wahrnehmung“ daher „Halluzinationen“ bzw. „Konfabulationen“
- Die Korrektheit des Textes muss **immer** überprüft werden
- Nachtrainiert, um nur politisch korrekte Aussagen zu liefern
- Mit RAG werden die Modelle wesentlich **vertrauenswürdiger**
- Überprüfen der **Quellen** (soferne angegeben)
- Multimodalität erweitert den „Horizont“ in Richtung größerer Glaubwürdigkeit



WHO AM I?

Ao.Univ.-Prof. am Institut für Artificial Intelligence
(ai.meduniwien.ac.at)

Curriculumdirektor für Medizinische Informatik

Pionier für Neuronale Netze in Österreich

Vizepräsident der ASAI (www.asai.ac.at)

Gründer der Siesta Group Schlafanalyse GmbH
(www.thesiestagroup.com)