SOL zu KI, LLM und Folgen fehlerhafter Nutzung

Irka Lohnitz

Erster Einstieg, Begriffserläuterungen, interessante Medien

Funktionsweise von Künstlicher Intelligenz (KI) und Sprachmodellen (LLM)

Künstliche Intelligenz (KI) und große Sprachmodelle (LLM) sind Technologien, die darauf abzielen, menschenähnliche Denk- und Sprachfähigkeiten zu simulieren.

Grundlagen der KI

- **Datenaufnahme**: KI benötigt große Mengen an Daten, um zu lernen. Diese Daten können Text, Bilder oder andere Informationen sein.
- **Lernverfahren**: KI-Systeme verwenden Algorithmen, um Muster in den Daten zu erkennen. Das häufigste Verfahren ist das maschinelle Lernen, bei dem Modelle mit Beispieldaten trainiert werden.

Was sind LLMs?

- **Training**: LLMs (z.B. GPT-3, DeepSeek) werden auf riesigen Textdatensätzen trainiert. Sie lernen, wie Wörter und Sätze zusammenhängen, um menschliche Sprache in Kontext zu setzen und zu generieren.
- **Transformer-Architektur**: Viele LLMs nutzen die Transformer-Architektur, die es ihnen ermöglicht, den Kontext von Wörtern in einem Satz besser zu verstehen. Dadurch können sie relevantere und kohärentere Antworten geben.

Nutzung von LLMs

- **Textgenerierung**: LLMs können Texte verfassen, Fragen beantworten oder sogar Geschichten erzählen, indem sie das Gelernte anwenden.
- **Kontextverständnis**: Sie speichern Informationen über Gespräche, wodurch sie auf vorherige Interaktionen eingehen können.

Herausforderungen

- **Bias und Fehler**: Aufgrund der Daten, auf denen sie trainiert wurden, können LLMs Vorurteile reproduzieren oder falsche Informationen liefern.
- **Interpretation**: Manchmal können sie komplexe Fragen oder Nuancen in der Sprache missverstehen.
- KI und LLM sind also leistungsstarke Werkzeuge, die durch das Lernen aus großen Datenmengen in der Lage sind, menschenähnliche Texte zu erstellen, aber sie haben auch ihre Grenzen.

KI => künstliche Intelligenz
AI => artificial intelligence
LLM => Large Language Model (großes Sprachmodell)

Zu den großen Sprachmodellen zählen die folgenden GPT-Modelle:

- OpenAI, z. B. GPT-3, die in ChatGPT und Microsoft Copilot verwendet werden
- Google, z. B. PaLM, Gemini und weitere Varianten
- Meta LLaMA-Sprachmodelle (Open-Source)
- Anthropics Claude und
- XAI Grok

Daneben gibt es auch leistungsfähige LLMs chinesischer Firmen wie diejenigen von Alibaba, DeepSeek, 01 AI und Zhipu AI.

Interessante Medien zum Thema KI auf die ich gestoßen bin (die ich selber noch erkunden möchte)

I Have No Mouth, and I Must Scream - Harlan Ellison, 1967
(Postapokalyptische, fiktive Horror-Kurzgeschichte, in welcher ein Supercomputer die Welt übernimmt. 1995 wurde darauf basierend mit Ellison auch ein Computerspiel entwickelt.)

Weapons of Math Destruction - Cathy O'Neil, 2016
(Nicht-fiktives Werk, welches sich mit Großdatenalgorithmen und deren Beitrag zur Vertiefung von bestehenden Ungleichheiten/Vorurteilen)

Dead Internet Theory
(Bots answering Bots online - und wie Soziale Medien dadurch noch unsozialer werden als sie eh schon sind.)

Schönen, kurzes (ca. 30min) Video von Florian Dalwigk zum Thema:
Wie KRIMINELLE künstliche Intelligenz nutzen! Vortrag bei der #IHK
https://www.youtube.com/watch?v=jaxkwYAM1JY

Geschichte der künstlichen Intelligenz

- ernsthafte Forschungen in die Thematik seit den 1950er Jahren
    - angeregt vor allem durch Alan Turings Aufsatz "Computer Machinery and Intelligence", Kapitel 3 "Elektronische und technische Grundlagen"

- formaler Startschuss der akademischen KI-Forschung 1956 in den USA
    - Dartmouth College in einem Sommer-Workshop

- frühes Scheitern an mangelnder Geschwindigkeit und Speicherkapazität der eingesetzten Computer
    - heute kein Problem mehr durch technische Fortschritte, dennoch wurde noch keine "starke KI" entwickelt, welche tatsächlich so denken kann wie ein Mensch

- zurzeit existiert nur "schwache KIs"
    - domänenspezifische KI
        - kein universell einsetzbares Äquivalent zum menschlichen Geist, sondern Algorithmen, die möglichst selbstständig spezifische Aufgaben lösen können
    - sind sehr vielseitig z.B. Bilderkennung, Stimmenerkennung, Erzeugen von Bild/Text/Ton, Brett-/Karten-/Videospiele, autonome Auskunftssysteme im Einzelhandel oder Onlineshops, teil- oder vollautonom fahrende Fahrzeuge, etc.

- aktueller KI-Boom
    - durch Chat-bots wie ChatGPT oder auf Zuruf Bild erzeugende KIs wie DALL-E (beide von der Firma OpenAI)

  → großer Zuwuchs von öffentlicher Aufmerksamkeit und Nutzung

- ChatGPT basiert auf dem Large Language Model (LLM)
    - besitzt bei der Version GPT-3 ca. 175 Millionen Parameter, und bei GPT-4 sogar über 1 Billionen Parameter, und wurde mit fast allen öffentlich im Internet verfügbaren Texten als Trainingsdaten gefüttert worden
- Antworten machen auf ersten Blick einen kompetenten Eindruck, sind aber nicht immer faktisch korrekt
    - z.B. Anfrage "Erfasse einen wissenschaftlichen Bericht zu folgendem Thema […]"
        - Ergebnis mit Gliederung, Einleitung etc. Sieht überzeugend aus, doch sind Quellenangaben oft frei erfunden
    - generierter Code
        - ist oft nach mehreren Durchläufen mit Hinweisen auf Problemen immer noch fehlerhaft
    - mathematisch-logische Fragestellungen
        - Ergebnisse können stark variieren, von korrekt bis grundlegend falsch vom Ansatz an

Weitere Problematiken mit KI

- in KI gefütterte Daten werden oft als weitere Trainingsdaten für das Model verwendet, auch fehlerhafte Ergebnisse, sowie persönliche oder vertraute Daten, was zu Datenschutzproblemen führen kann

- die Trainingsdaten von KIs, welche Bilder generieren, sind oft auch geschützte Werke von Künstler*innen die keine Einwilligung für diese Art von Nutzung ausgesprochen haben, was zu Problemen mit Urheberrechten führen kann
    - dies gilt auch für geschriebene Werke und Musik, wodurch einige Künstler*innen ihre Werke nicht mehr öffentlich zur Anschauung stellen, damit diese nicht ungewollte entwendet werden

Wie lernt KI?

- überwachtes Lernen
    - Trainingsdaten, bei denen gewünschte Schlussfolgerungen bekannt ist, bis Daten mit noch unbekannten Zuordnungen vom Algorithmus interpretiert und eingeordnet werden können

- unüberwachtes Lernen
    - keine Beispieldaten, Algorithmus muss z.B. Kategorien selbst einteilen durch ähnliche Merkmale in den gegebenen Daten

- verstärktes Lernen
    - mit Bewertungsschema z.B. 5 Sterne bei einer richtigen Antwort des Algorithmus, 1 Stern bei einer falschen
    - Schritt für Schritt bis der Algorithmus in die richtige Richtung gelenkt wird
        - kann durch die menschliche Bewertung zu einem "Yes-Man"-Effekt führen, weil das Programm merkt bei positiver Zurede der Aufgabe "belohnt" zu werden, egal ob die gegebene Antwort faktisch richtig oder falsch ist

Quelle:
- IT-Handbuch 2023 für Fachinformatiker*innen Der Ausbildungsbegleiter, 11., aktualisierte und überarbeitete Auflage, Sascha Kersken

Beispielfälle von fehlerhafter KI-Nutzung mit gravierenden Folgen

Schulwesen

- UK, 2020:
  - während der Coronapandemie fielen fast alle Prüfungen aus
    - Schüler*innen wollte man dennoch ihre Abschlussnoten vergeben, damit diese das Jahr nicht wiederholen müssen
  - für Benotung wurde von Ofqual (Office of Qualifications and Examinations Regulation) ein Algorithmus für diese Zwecke entwickelt und eingesetzt
  - nach den Berechnungen des Algorithmus und deren Bekanntgabe, wurde allerdings von Schüler*innen festgestellt, dass die Ergebnisse oft schlechter als die Prognosen ihrer Lehrer*innen waren
    - ca. 36% der A-Level Benotungen waren um eine Note verschlechtert, 3% sogar um zwei
    - im Vergleich war es bei Studienanfänger*innen noch verheerender, 79% lagen unter ihren erwarteten Schnitt
  - ebenso wurde festgestellt, dass private Schulen im Schnitt besser als öffentliche, und besonders Schulen in schwächeren sozialökonomischen Bezirken, abgeschnitten haben
    - der Algorithmus soll sich auf Ergebnisse der Vorjahre bezogen haben, was die Leistungen der Schüler in schwächeren Regionen besonders untergraben hat
  - Resultat waren große Protestaktionen, welche letztendlich dazu geführt haben, dass die Abschussnoten mit Hilfe der Lehrerprognosen korrigiert wurden
    - dadurch konnten z.B. ca. 15,000 Schüler*innen ihren gewünschten Berufsweg weiterverfolgen, welcher davor durch die Algorithmus-Ergebnisse verbaut wurde

Quelle:

- https://en.wikipedia.org/wiki/2020_United_Kingdom_school_exam_grading_controversy
- https://en.wikipedia.org/wiki/Ofqual_exam_results_algorithm

Bewerbungsverfahren

- USA:
  - es werden vermehrt KI-Systeme für Bewerbungsverfahren eingesetzt
    - z.B. Systeme wie HireVue analysieren Sprachmuster, Tonfall oder Gesichtsbewegungen
      - Problematik hierbei ist, dass das System abweichende Sprechweisen oder Bewegungen als "fehlerhaft/unerwünscht" einordnet, was automatisch zu einem Bias und Diskrimination von Menschen mit Behinderungen führt
        (z.B. wenn die Person ein Stottern haben, Ticks, Schwierigkeiten mit Augenkontakt durch evtl. autistische Hintergründe und ähnliche ununterdrückbare bzw. angeborene Eigenschaften, unabhängig von den eigentlichen Qualifikationen der Person)

        - der Bias solcher Systeme "füttert" sich nicht nur mit den schon vorhandenen Vorurteilen der Entwickler/Arbeitgeber, ob unbewusst implementiert oder nicht, sondern auch mit seinen eigenen Ergebnissen, wodurch sich nicht nur die schon bestehenden Vorurteile verstärken, sondern teilweise auch neue Diskriminationsmuster schafft

    - Amazon musste z.B. sein KI-Bewerbungsverfahren (2014-2017) einstellen, da es Bewerbungen, besonders in z.B. IT-Bereichen, nicht geschlechtsneutral bewertet hat
        - dem Algorithmus wurden ältere Bewerbungen zu Verfügung gestellt, um daraus die besten Neubewerbungen herauszufiltern
            - da es sich aber um einen von hauptsächlich Männern dominierter Bereich handelt, wurden Frauen rausgefiltert, da die meisten erfolgreichen Bewerbungen quantitativ von Männern waren, wodurch Frauen automatisch benachteiligt wurden, selbst bei gleichen oder sogar besseren Qualifikationen

Quelle:

- https://thehill.com/opinion/technology/4576649-ai-is-causing-massive-hiring-discrimination-based-on-disability/
- https://www.cnbc.com/2018/10/10/amazon-scraps-a-secret-ai-recruiting-tool-that-showed-bias-against-women.html

Sozialleistungen

- Niederlande:
    - niederländische Regierung musste 2021 ein Bußgeld von 2,75 Millionen Euro zahlen wegen massiver Verstöße gegen die Datenschutzgrundverordnung
    - im Rahmen der Toeslagenaffaire (Kindergeldaffäre) in den 2010er Jahren hat die Steuerbehörde, durch Nutzung eines Algorithmus, Informationen über die Nationalität und Lebensumstände der Betroffenen genutzt, um diese fälschlicherweise des Kindergeldbetrugs zu beschuldigen
        - zehntausende Eltern waren betroffen, von welchen hohe Rückzahlungen des Kindergeldes gefordert wurden
        - selbst kleine Formfehler führten zu erheblichen Nachforderungen und verursachten Jahre lang stigmatisierende Betrugsermittlungen
        - viele Familien, die auf staatliche Unterstützung angewiesen sind, wurden dadurch zu Unrecht in den Ruin getrieben
    - Informationen über Staatsangehörigkeit wurden zu Unrecht im Rahmen von Betrugsbekämpfung als Verdachtsmarker verwendet
        - im automatisierten Risikoprüfungssystem galt eine nicht-niederländische Nationalität als Risikofaktor
    - 2020 wurde die systematische Diskriminierung erst nach Jahren entdeckt
        - 2021 folgten Rückzahlungen und Entschädigungen an die Opfer, welche allerdings jahrelang um Aufklärung und Anerkennung kämpfen mussten

Quelle:

- https://netzpolitik.org/2021/kindergeldaffaere-niederlande-zahlen-millionenstrafe-wegen-datendiskriminierung/