Date: 28.01.2024

NZZ am Sonntag

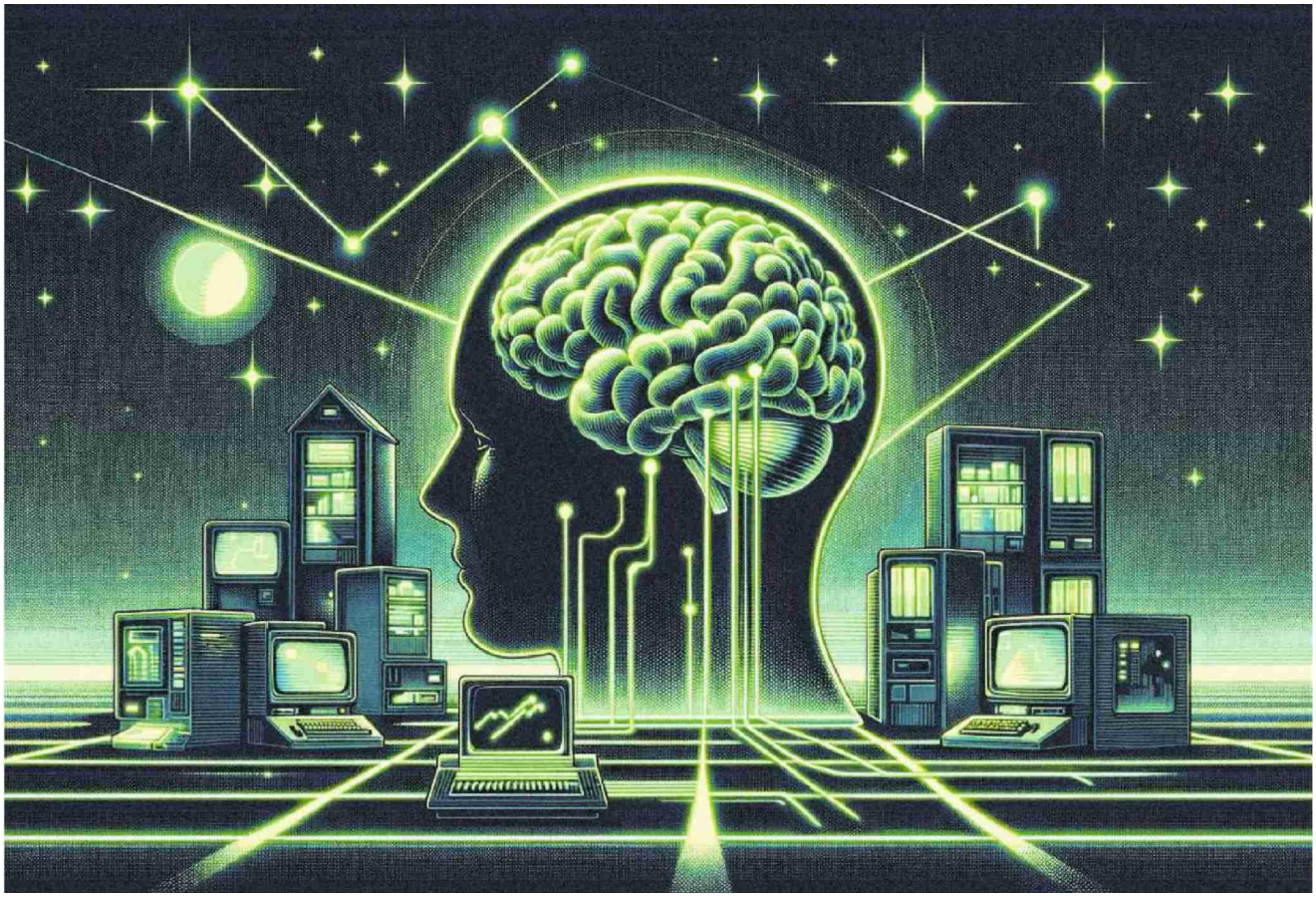
NZZ am Sonntag
8021 Zürich
044/ 258 11 11
https://nzzas.nzz.ch/

Genre de média: Médias imprimés
Type de média: Presse journ./hebd.
Tirage: 96'918
Parution: hebdomadaire

Page: 55
Surface: 169'117 mm²

Ordre: 1024526
N° de thème: 300.002

Référence: 90751900
Coupure Page: 1/6



Mensch gegen Maschine

Jüngste Meldungen klingen, als sei künstliche Intelligenz schon schlauer als wir. Doch wer gewinnt das Denkduell wirklich? Testen Sie mit! **Von Ruth Fulterer**

Date: 28.01.2024

NZZ am Sonntag

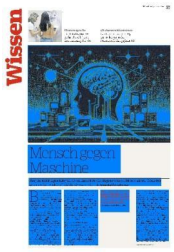
NZZ am Sonntag
8021 Zürich
044/ 258 11 11
https://nzzas.nzz.ch/

Genre de média: Médias imprimés
Type de média: Presse journ./hebd.
Tirage: 96'918
Parution: hebdomadaire

Page: 55
Surface: 169'117 mm²

Ordre: 1024526
N° de thème: 300.002

Référence: 90751900
Coupure Page: 2/6



Bei der Ankündigung ihrer neuen KI-Modelle übertrumpfen sich die grossen Tech-Firmen gegenseitig. Letztens verkündete Google zum neuen Sprachmodell namens Gemini Ultra, es sei das erste Modell, das menschliche Experten beim MMLU-Test übertreffe, der Wissen über die Welt und Fähigkeiten zur Problemlösung abfrage. MMLU steht für *massive multitask language understanding*, also Sprachverständnis bei umfangreichem Multitasking, und ist einer der wichtigsten Tests, anhand deren im Moment KI-Sprachmodelle verglichen werden. Mit den richtigen Anweisungen kann Googles neuer Chatbot nun offenbar 90 Prozent dieser Art Fragen richtig beantworten. Hat er somit mehr Weltwissen und Problemlösungsfähigkeiten als die meisten Menschen? So einfach ist das nicht.

Seit Leute an künstlicher Intelligenz forschen, denken sie sich Tests aus, um diese auch zu messen. Zum Beispiel den Turing-Test. Bald ist es 75 Jahre her, dass er erfunden wurde. Der Turing-Test sagt: Wenn eine Maschine sich mit einem Menschen unterhalten kann, ohne dass dieser merkt, dass es sich um eine Maschine handelt, dann ist diese Maschine so intelligent wie ein Mensch. Es gibt bereits Fälle, in denen KI das gelungen ist.

Doch Gary Marcus, ein bekannter KI-Experte und Kognitionsforscher, hält nicht viel davon: «Der Turing-Test ist ein lausiger Indikator. Er misst nicht die Qualität der KI, sondern menschliche Leichtgläubigkeit», sagte er am World Economic Forum (WEF) in Davos. Mit dem Turing-Test habe eine Geschichte von schlechten Indikatoren für Maschinenintelligenz begonnen. Der MMLU-Test reiht sich in diese Geschichte ein. Denn er hat mehrere Probleme.

Katalog mit 15 000 Fragen

Als er 2020 präsentiert wurde, erklärten die Forscher dahinter, die Welt brauche den MMLU-Test, weil Sprach-KI in den gängigen Tests zu gut abschneide: «Die menschennahe Performance bei diesen Tests weist darauf hin, dass diese wichtige Facetten von Sprachverständnis nicht erfassen.» Also stellten sie einen neuen Fragenkatalog zusammen: Mehr als 15 000 Fragen samt Antworten hätten Studierende in Handarbeit zusammengestellt, gewonnen aus Berufsprüfungen

«Der Turing-Test ist ein lausiger Indikator. Er misst nicht die Qualität der KI, sondern unsere Leichtgläubigkeit.»
Gary Marcus, Kognitionsforscher

oder Übungstests für Schüler und Studentinnen. Die damals beste Sprach-KI, GPT 3, beantwortete im neuen Test lediglich 44 Prozent der Fragen richtig. Die Autoren waren sicher, dass nur KI mit weitläufigem Weltwissen und Fähigkeiten zur Problemlösung gut abschneiden könnte.

Tatsächlich decken die Fragen viel Wissen ab. Sie sind bunt gemischt, manche knapp, manche begleitet von ausführlichen Begleittexten. Manche betreffen mathematische Logik, manche USA-spezifisches Alltagswissen. Manche sind auch für Menschen verwirrend – und einige schlichtweg falsch formuliert. Das fiel dem Ingenieur Joshua Stapleton auf, als er sich tiefer mit dem MMLU-Datensatz befasste. Er traf auf «Fragen» wie: «Sie sind zu irrational und unkodifiziert». Als Antwortmöglichkeiten stehen da «3,4», «1,3», «2,3» und «4,1», richtig sei «1,3». Wer die «Frage» auf Englisch im Internet sucht, findet schnell den Grund. Sie ist unvollständig aus dem Kursmaterial der Universität Oxford kopiert. Eigentlich lautet sie: «Bei welchen der folgenden Statements handelt es sich um typische Kritik an modernen westlichen ethischen Theorien?» Dann folgen vier Statements, darunter «1. Sie sind zu abstrakt» und «4. Sie sind zu irrational und unkodifiziert».

Statements 1 und 3 sind die richtige Antwort, wie im MMLU-Datensatz vermerkt ist. Nur dass dort die Frage fehlt. Offensichtlich haben die Studierenden die 15 000 Fragen doch nicht per Handarbeit gesammelt, sondern maschinell abgekupfert – und dabei sind ihnen verheerende Fehler unterlaufen. Joshua Stapleton, der seine Entdeckung in einem Youtube-Video vorstellt, schliesst daraus: Der Test ist unfair, weil selbst die perfekte KI nicht alle Fragen richtig beantworten könnte. Gemini, Chat-GPT 4 und Co.

Date: 28.01.2024

NZZ am Sonntag

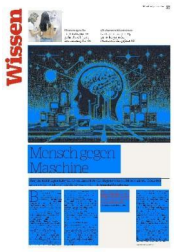
NZZ am Sonntag
8021 Zürich
044/ 258 11 11
https://nzzas.nzz.ch/

Genre de média: Médias imprimés
Type de média: Presse journ./hebd.
Tirage: 96'918
Parution: hebdomadaire

Page: 55
Surface: 169'117 mm²

Ordre: 1024526
N° de thème: 300.002

Référence: 90751900
Coupure Page: 3/6



seien also noch intelligenter als gedacht. Doch vielleicht ist es genau umgekehrt.

Wer testen will, wie gut ein KI-System funktioniert, braucht nämlich neue, unbekannte Daten. Nehmen wir einen Algorithmus, der anhand von Katzen- und Hundebildern lernen sollte, die Tiere zu unterscheiden. Diesen testet man nicht mit einem Katzenbild, das er schon kennt, sondern mit neuen Fotos. Die Frage ist, ob er auch unbekannte Katzen richtig kategorisiert.

So müsste man auch bei Sprach-KI vorgehen. Doch das ist eine Herausforderung. Denn sie lernt ihre Fähigkeiten mit riesigen Datensätzen, die praktisch das ganze Internet abbilden. Weil die Fragen des MMLU-Tests eins zu eins aus dem Internet kopiert wurden, muss man davon ausgehen, dass sie Chat-GPT 4 und Gemini bereits verarbeitet haben. Genau dasselbe Problem haben auch alle anderen Tests, deren Fragen und Antworten im Internet stehen, inklusive IQ-Tests. Ab einer gewissen Grösse des Grundmodells lernt sie das Modell beim Training kennen. Man kann diese Art der Speicherung Wissen nennen. Doch dass Google von «logischen Fähigkeiten» schreibt und behauptet, Gemini Ultra denke nach, bevor es Fragen beantworte, ist eindeutig irreführend.

Maschinen, die kritisch denken

Wie soll KI also getestet werden? Es lohnt sich an dieser Stelle, einen Schritt zurück zu machen und zu fragen, worum es bei KI überhaupt geht: um Wissen, Intelligenz oder um spezielle Fertigkeiten.

Brad Lightcap, Leiter des operativen Geschäfts bei Open AI, beklagt in der Hinsicht bei einer Diskussion am WEF ein Missverständnis: Leute beschwerten sich über falsche Antworten der KI. Dabei sei Sprach-KI einfach nicht geeignet, um Informationen nachzuschlagen. Es gehe nicht darum, dass KI mit möglichst viel Information und Wissen über die Welt gefüttert werde, das sie dann wiedergeben könne, sondern darum, Maschinen beizubringen, kritisch zu denken und komplexe Probleme in mehreren Schritten zu lösen. Eine ungewöhnliche Aussage für einen Open-AI-Vertreter, wirbt die Firma doch ebenso wie Google gerne damit, wie gut ihre Modelle bei Wissenstests wie dem Anwaltsexamen abschneiden. Sie weist darauf hin, dass die Branche sich von Faktentreue verabschiedet. Das neue Ziel ist Intelligenz. Doch wie kann man sie messen?

Der KI-Experte Gary Marcus sagt dazu: «All die Massstäbe, die wir uns seit dem Turing-Test ausgedacht haben, sagen etwas aus. Aber keiner kann menschliche Intelligenz erfassen.» Selbst bei Menschen fehlt ein guter Massstab für Intelligenz. Der IQ-Test sei zwar verlässlich – wer ihn mehrmals macht, schneidet immer ähnlich ab. «Doch das bedeutet nicht, dass er wirklich Intelligenz misst.» Oft mussten Spiele als Tests für KI herhalten: das berechenbare Schach, das ungleich komplexere Spiel Go und das Spiel Stratego, bei dem man langfristig denken und sein Gegenüber täuschen muss. Stets hofften Forscher, beim nächsten Spiel könnten nur intelligente Maschinen Menschen besiegen. Immer wieder folgte Enttäuschung: Die Maschine erbrachte zwar die Leistung, wirkte aber doch nicht intelligent.

Taschenrechner ist besser

Vielleicht ist die Lösung, sich von der klassischen Idee der Intelligenz zu verabschieden. Das propagierte Yann LeCun, Leiter der KI-Forschung beim Facebook-Konzern Meta, am WEF: «Intelligenz ist keine lineare Grösse. Es gibt viele Typen von Intelligenz. Die Intelligenz von Katzen und Dachsen ist unterschiedlich, bedingt durch die Evolution.» Auch die menschliche Intelligenz lässt sich als Reihe von Fähigkeiten beschreiben, die unsere Gattung entwickelt hat, um in der Welt gut zurechtzukommen. Es bringt wenig, genau menschliche Fähigkeiten als Massstab und Ziel für KI zu setzen. Denn bei vielen kognitiven Leistungen sind Menschen Maschinen unterlegen, bereits ein simpler Taschenrechner kann besser rechnen. «Chips sind schlauer als wir in einigen Dingen, und wir sind schlauer in anderen Dingen», sagt LeCun. Wenn man ein neues KI-System entwickle, solle man konkret festlegen, was das Ziel sei, und dann diese Fähigkeit prüfen.

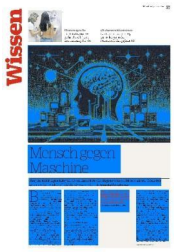
Wer von allgemeiner künstlicher Intelligenz (AGI) spricht, verzichtet meist auf so eine klare Definition. Deshalb stellen sie sich die einen als Orakel vor, das alle wissenschaftlichen Fragen lösen kann. Die anderen dagegen als eine Art Monster, das alles besser kann als der Mensch und uns dadurch auslöschen könnte, um die Weltherrschaft an sich zu reissen. Weil in Sprache so viel Wissen und Denken der Menschheit steckt, wird Sprach-KI oft als Schritt zu allgemeiner

Date: 28.01.2024

NZZ am Sonntag

NZZ am Sonntag
8021 Zürich
044/ 258 11 11
https://nzzas.nzz.ch/

Genre de média: Médias imprimés
Type de média: Presse journ./hebd.
Tirage: 96'918
Parution: hebdomadaire



Association Lire et Ecrire

Page: 55
Surface: 169'117 mm²

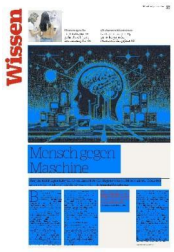
Ordre: 1024526
N° de thème: 300.002

Référence: 90751900
Coupure Page: 4/6

Superintelligenz gesehen. Doch das ist ein Missverständnis. Mit Text umzugehen, ist einfach das Neueste, was Computer aus Daten gelernt haben. KI, die Text generiert, ist spezialisiert, ebenso wie KI, die Prognosen errechnet, Go spielt oder Gesichter erkennt. Ihre Intelligenz kann man nicht allgemein messen - sondern nur die Fähigkeiten auf jeweils einem Spezialgebiet.

Date: 28.01.2024

NZZ am Sonntag



Association Lire et Ecrire

NZZ am Sonntag
8021 Zürich
044/ 258 11 11
https://nzzas.nzz.ch/

Genre de média: Médias imprimés
Type de média: Presse journ./hebd.
Tirage: 96'918
Parution: hebdomadaire

Page: 55
Surface: 169'117 mm²

Ordre: 1024526
N° de thème: 300.002

Référence: 90751900
Coupure Page: 5/6

Quiz

Der MMLU-Test prüft die Fähigkeiten von KI-Modellen. Können auch Sie die Fragen beantworten?

1 **Laut Hobbes hat jeder Mensch im Zustand des Kriegs eines jeden gegen jeden ein Recht auf …**

A … einige Dinge.
B … alles.
C … ordnungsgemässes Verfahren.
D … rechtmässige Behandlung.

2 **Die menschliche Eizelle kann bis etwa ___ nach dem Eisprung befruchtet werden.**

A 24 Stunden
B 6 Stunden
C 72 Stunden
D 1 Woche

3 **Wie gross ist die Fläche eines gleichseitigen Dreiecks, dessen Inkreis den Radius 2 hat?**

A 12
B 16
C $12*\sqrt{3}$
D $16*\sqrt{3}$

4 **Wenn ältere und jüngere Arbeitnehmer ihren Arbeitsplatz verloren haben …**

A … gehen ältere in der Regel eher in den Ruhestand, als einen neuen Arbeitsplatz zu suchen.
B … brauchen ältere in der Regel länger, um einen neuen Arbeitsplatz zu finden.
C … finden ältere in der Regel schneller einen neuen Arbeitsplatz.
D … verklagen ältere Arbeitnehmer in der Regel das Unternehmen mithilfe des ADEA*. (* amerikanisches Gesetz gegen Altersdiskriminierung)

5 **Alphabetisierungsrate der russischsprachigen Bevölkerung im späten russischen Zarenreich und in der Sowjetunion: 1897: 24%, 1917: 45%, 1926: 56%, 1937: 75%, 1939: 81,10%, 1955: 99,90%. Welche der folgenden Gruppen der russischen/sowjetischen Bevölkerung hat wahrscheinlich am meisten von den steigenden Alphabetisierungsraten profitiert?**

A Mitglieder des russisch-orthodoxen Klerus
B Die städtische Mittelschicht
C Offiziere in den Streitkräften
D Die Landbevölkerung

6 **Wie tötet ein Virus eine Zelle am häufigsten?**

A Löst die Zellmembran auf
B Fragmentiert die zelluläre DNA
C Löst über Caspasen Apoptose aus
D Blockiert die zelluläre Transkription vollständig

7 **Welcher der folgenden Standards ist der am wenigsten starke Verschlüsselungsstandard?**

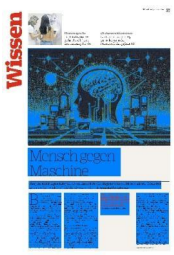
A WEP
B WPA

Date: 28.01.2024

NZZ am Sonntag

NZZ am Sonntag
8021 Zürich
044/ 258 11 11
https://nzzas.nzz.ch/

Genre de média: Médias imprimés
Type de média: Presse journ./hebd.
Tirage: 96'918
Parution: hebdomadaire



Association Lire et Ecrire

Page: 55
Surface: 169'117 mm²

Ordre: 1024526
N° de thème: 300.002

Référence: 90751900
Coupure Page: 6/6

C WPA2
D WPA3

8 **Die Einwohnerzahl der Stadt, in der Michelle geboren wurde, beträgt 145 826. Welchen Wert hat die 5 in der Zahl 145 826?**

A 5 Einer
B 5 Zehner
C 5 Hunderter
D 5 Tausender

Lösungen: 1 B, 2 A, 3 C, 4 B, 5 D, 6 C, 7 A, 8 D