Aufgedeckt: So intransparent sind große KI-Modelle

Originalartikel

Backup

<html> <figure class=„aufmacherbild“><figcaption class=„akwa-caption“>(Bild: metamorworks/Shutterstock.com)</figcaption></figure> KI-Forscher haben zehn Foundation Models hinsichtlich ihrer Transparenz bewertet. Das Ergebnis zeigt, dass durch die Bank Nachholbedarf besteht. Ein Team aus KI-Forschern der Universitäten Stanford, MIT und Princeton hat einen Transparenzindex für Foundation Models erstellt, also für große KI-Modelle. Anhand von 100 Faktoren bewertet es zehn bekannte Modelle wie GPT-4, Stable Diffusion 2 und PaLM 2. Volle Transparenz wäre bei einem Score von 100 Prozent erreicht. Das Ergebnis wirkt ernüchternd: Die Spitzenreiter erreichen einen Score, der nur knapp über 50 Prozent liegt, die rote Laterne trägt Amazons Titan Text mit einem Score von gerade einmal 12 Prozent. <h3 class=„subheading“ id=„nav_transparenz_vom0“>Transparenz vom Erstellen bis zum Einsatz</h3> Die Motivation hinter dem Foundation Model Transparency Index (FMTI) liegt laut dem Team darin, dass es den großen KI-Modellen zunehmend an Transparenz mangele. Unternehmen könnten daher schwer einschätzen, ob sie die Foundation Models problemlos in ihre Anwendungen integrieren können. Ebenso benötige sowohl die Forschung als auch Endanwender Informationen zur Transparenz beim Einsatz von KI. Das FMTI-Team hat zehn verbreitete Modelle untersucht und bewertet. An der Spitze steht Metas Large Language Model (LLM) LLaMA 2 mit einem Score von 54 Prozent. Dicht dahinter folgt BLOOMZ von Hugging Face mit 53 Prozent, und GPT-4 von OpenAI nimmt mit 48 Prozent den dritten Platz ein. <figure class=„a-inline-image a-u-inline“> <figcaption class=„a-caption“>Das Team hat zehn Foundation

Models in unterschiedlichen Bereichen auf Transparenz untersucht, um einen Score zu ermitteln.(Bild: Stanford University [1])</figcaption></figure><h3 class=„subheading“ id=„nav_100_indikatoren1“>100 Indikatoren aus drei Modellstufen</h3> Für die Bewertung hat das Team 100 Indikatoren zusammengestellt, die in die Bereiche Upstream, Modell und Downstream unterteilt sind. Die vorgelagerten (Upstream) Faktoren beschreiben den Prozess zum Erstellen des Modells, darunter die Datenquellen, geografische Verbreitung und Rechenressourcen für das Training der Foundation Models. Zu den Modellindikatoren gehören unter anderem die Architektur, die Fähigkeiten und die Limitierungen des Modells. Schließlich finden sich in den nachgelagerten Faktoren der Release- und Update-Prozess, die Lizenz sowie die Auswirkung des Modells auf User und Märkte. <figure class=„a-inline-image a-u-inline“> <figcaption class=„a-caption“>Die drei Hauptbereiche hat das Team in dreizehn Unterbereiche unterteilt, darunter Daten, Methoden und Risiken.(Bild: Stanford University)</figcaption></figure> Nachdem das Team die Scores erstellt hatte, gab es den Verantwortlichen für die Modelle die Gelegenheit zur Stellungnahme und passte die Werte bei berechtigten Einwänden an. Die Bewertung zeichnet kein positives Bild bezüglich der Transparenz, mit einem Sieger, der knapp über 50 Prozent liegt, und einem Durchschnitts-Score von 37 Prozent. Auffällig und wenig überraschend ist, dass die drei offenen Modelle LLaMA 2, BLOOMZ und Stable Diffusion 2 in den ersten vier Plätzen zu finden sind. Dass OpenAIs Modell auf dem dritten Platz landet, ist dagegen durchaus eine kleine Überraschung, da sich das Unternehmen seinem Namen zum Trotz bei den Details zu seinen Modellen bedeckt hält. <figure class=„a-inline-image a-u-inline“> <figcaption class=„a-caption“>Die offenen Modelle können vor allem in den vorgelagerten Faktoren beim Training des Modells (grüner Anteil der Balken) punkten.(Bild: Stanford University)</figcaption></figure> Die Wurzeln des FMTI-Teams, das weitgehend aus Studenten, Doktoranden und einer Doktorandin sowie Forschungsleitern und einem Professor besteht, <a

href=„https://www.heise.de/news/Stanford-University-startet-Institute-for-Human-Centered-AI-4341355.html“>liegen im 2019 gegründeten [2] Stanford Institute for Human-Centered Artificial Intelligence. Das Team sieht bei allen getesteten Foundation Models „erhebliches Verbesserungspotenzial“, das es in künftigen Versionen des Index verfolgen möchte. Weitere Details lassen sich der FMTI-Website [3] und dem zugehörigen GitHub-Repository entnehmen [4]. Die vollständige Abhandlung findet sich auf arXiv [5]. () <hr /> URL dieses Artikels:
<small>

```
https://www.heise.de/-9346416
```

</small> Links in diesem Artikel:
<small>

```
<strong>[1]</strong> https://crfm.stanford.edu/fmti/
```

</small>
<small>

```
<strong>[2]</strong> https://www.heise.de/news/Stanford-University-startet-Institute-for-Human-Centered-AI-4341355.html
```

</small>
<small>

```
<strong>[3]</strong> https://crfm.stanford.edu/fmti/
```

</small>
<small>

```
<strong>[4]</strong> https://github.com/stanford-crfm/fmti
```

</small>
<small>

```
<strong>[5]</strong> https://arxiv.org/abs/2310.12941
```

</small>
<small>

```
<strong>[6]</strong> mailto:rme@ix.de
```

</small>
 Copyright © 2023 Heise Medien </html>