Workshop:
Offenheit von KI-Systemen

Gesellschaftliche Auswirkungen und Mitgestaltungsmöglichkeiten

AUTOR:IN
Axel Dürkop

VERÖFFENTLICHUNGSDATUM
Mittwoch, 1. November 2023

ZUSAMMENFASSUNG
In diesem Workshop soll es darum gehen, Potenziale und Herausforderungen offener KI-Systeme zu diskutieren. Zunächst bestimmen die Teilnehmenden, was sie unter Offenheit verstehen wollen. Nach einer Würdigung der gemeinsam erarbeiteten Ergebnisse lernen sie ein Framework kennen, um KI-Systeme auf sechs implementierte Werte hin zu untersuchen. In einem nächsten Schritt wird die Plattform Hugging Face vorgestellt, die versucht, einige der genannten Werte aus dem Framework umzusetzen. An einer Grafik zu öffentlich zugänglichen Large Language Models (LLMs) diskutiert die Gruppe, ob damit ihre Anforderungen an Offenheit eingelöst werden. Im vorletzten Schritt wird das Verfahren des Fine-tunings erläutert und die Möglichkeiten beschrieben, LLMs auf eigene Anforderungen hin zu trainieren. Abschließend sammelt die Gruppe User Storys aus einer pädagogischen Perspektive und mit verschiedenen Rollen, um den aktuellen KI-Produkten eigene Vorstellungen von KI-Systemen im Bildungszusammenhang entgegenzusetzen.

Was wollen wir unter Offenheit verstehen?

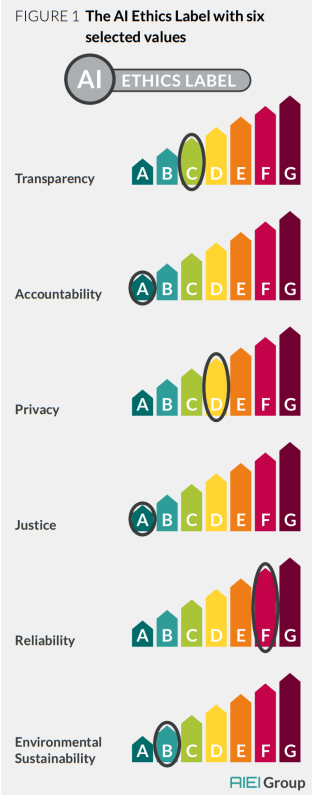
Was ist Offenheit?

Die folgenden Werte, die mit Offenheit verbunden werden, werden der Gruppe vorgestellt. Anschließend erfolgt die Bitte, diese zu diskutieren und wenn möglich das Verständnis von Offenheit weiter auszuarbeiten.

- Transparenz
- Partizipation
- Nachvollziehbarkeit
- Zugang
- Gerechtigkeit
- Privatheit

Welche ethischen Ansprüche haben wir an KI-Systeme?

Zentrale Werte von KI-Systemen

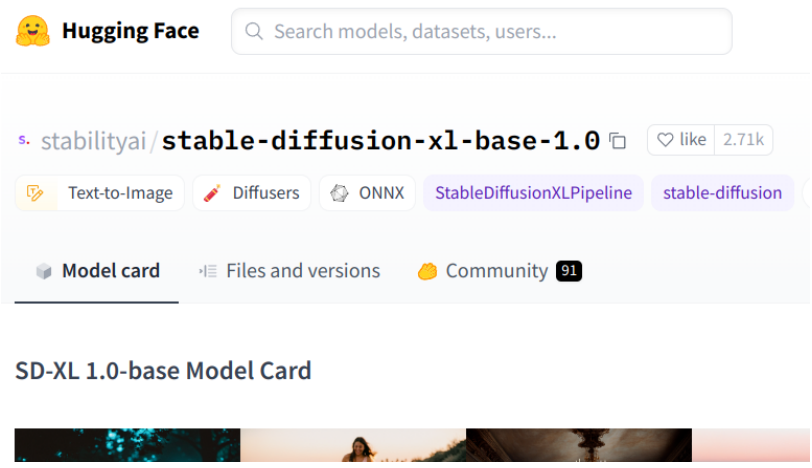


- **Transparenz** (*transparency*)
- **Rechenschaft** (*accountability*)
- **Datenschutz/Privatsphäre** (*privacy*)
- **Gerechtigkeit** (*justice*)
- **Zuverlässigkeit** (*reliability*)
- **Ökologische Nachhaltigkeit** (*environmental sustainability*)

"AI Ethics Label. Aus: From Principles to Practice. An interdisciplinary framework to operationalise AI ethics", Hallensleben et al. (2020, S. 13f.)

Wodurch zeichnen sich offene KI-Systeme aus?

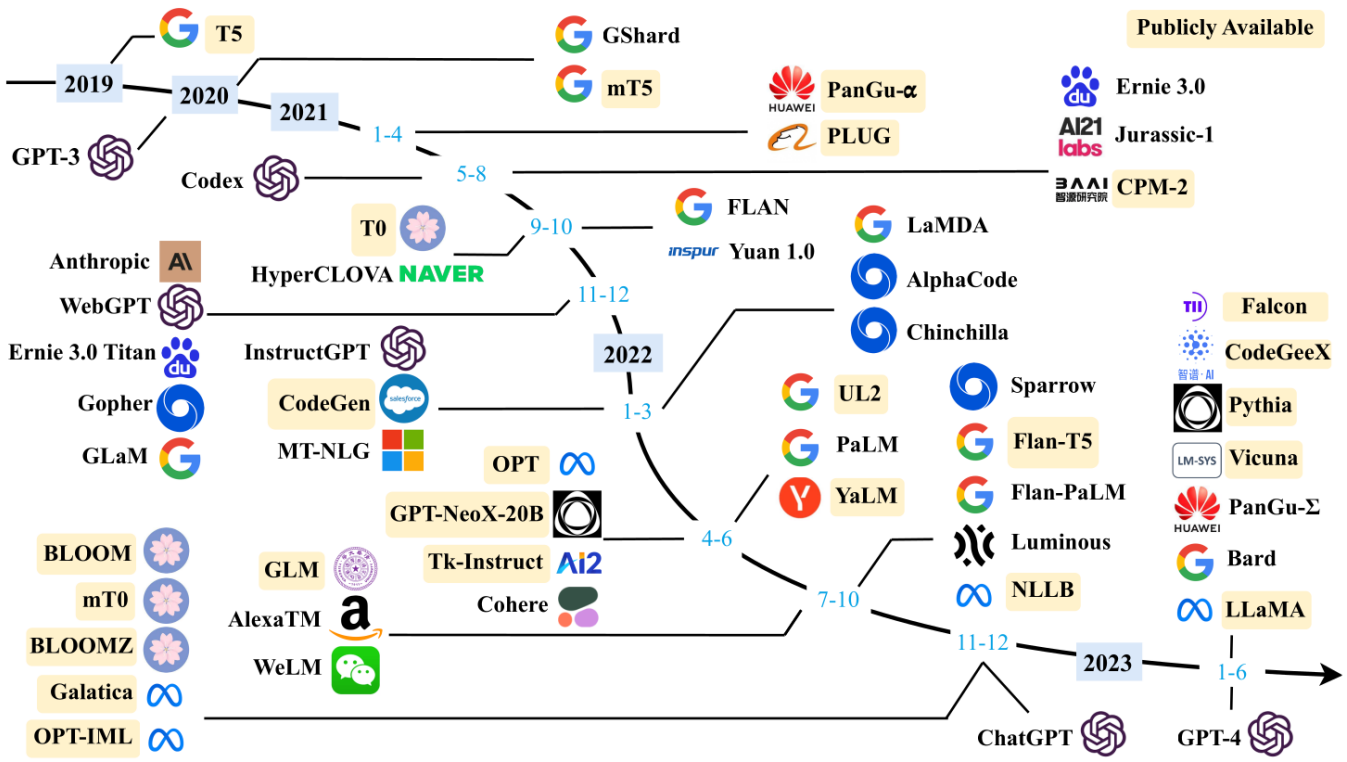
Einlösung der genannten Werte?



- Model Cards versuchen, Transparenz zu schaffen
- Trainingsdaten liegen offen und sind dokumentiert
- Modelle sind zur Nachnutzung freigegeben
- Quellcode für die Anwendungen ist frei verfügbar

Quelle: Screenshot von The Hugging Face

Öffentlich zugängliche LLMs = Offenheit?

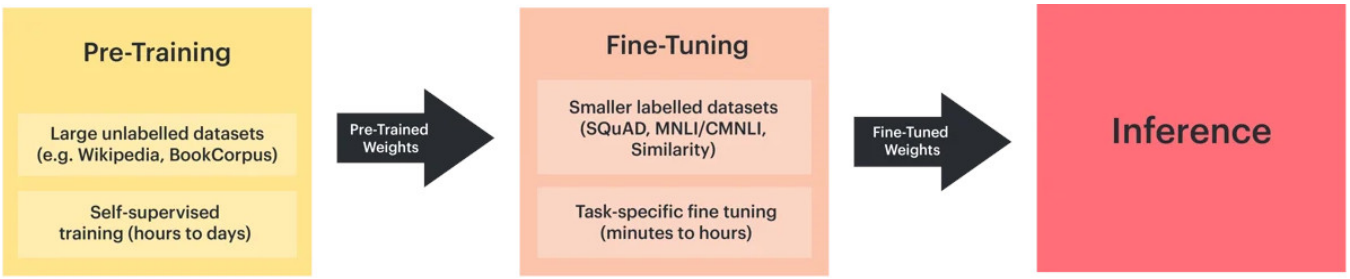


Gelb gekennzeichnet: offen zugängliche LLM. Quelle: Zhao et al. (2023, S. 7)

Fine-tuning: Der Schlüssel zu eigenen Anwendungen

Was ist Fine-tuning?

Beim Fine-tuning werden große Basismodelle mit kleineren aufgabenspezifischen Datensätzen trainiert.



Pre-training and fine-tuning BERT. Quelle: Chen & Brown (2021)

Welche Anwendungen wünschen wir uns?

User Storys

Aufbau einer **User Story** nach Cohn (2010):

> Als [1] möchte ich [2], damit/um/weil [3].

1. Nutzerrolle
2. Anforderung
3. Grund für die Anforderung

User Storys orientieren sich am gewünschten **Outcome**.

> Als Schülerin möchte ich mit dem Buch “Die Welle” chatten, weil so die Auseinandersetzung mit dem Text einfacher ist.

An welche Akteur*innen richten wir uns?

Die Gruppe sammelt Akteur*innen, an die die User Storys gerichtet werden können.

Akteur*innen im Bereich KI-Systeme

- IT-Konzerne
- Schulen
- Verlage
- …

Was müssen wir wissen?

Sofern die Zeit es zulässt, können abschließend noch Kompetenzen gesammelt werden, die die Rollen in den User Storys für die Umsetzung der Vorstellungen von offenen KI-Systemen haben müssen.

Kontakt

> https://axel-duerkop.de
> me@axel-duerkop.de
> @xldrkp@scholar.social

Nachnutzung

Acknowledgments

Der Workshop, für den dieses Skript erstellt wurde, fand am 21. September 2023 im Rahmen des Forums Medienethik in Soltau statt. Die Entwicklung des Workshops wurde finanziert durch das Niedersächsische Landesinstitut für schulische Qualitätsentwicklung (NLQ).

Referenzen

Chen, J., & Brown, P. (2021, Mai 6). *BERT-Large Training on the IPU Explained* [Unternehmenswebsite]. Graphcore. https://www.graphcore.ai/posts/bert-large-training-on-the-ipu-explained

Cohn, M. (2010). *User Stories* (M. Hesse-Hujber, Übers.). mitp.

Hallensleben, S., Hustedt, C., Fetic, L., Fleischer, T., Grünke, P., Hagendorff, T., Hauer, M., Hauschke, A., Heesen, J., Herrmann, M., Hillerbrand, R., Hubig, C., Kaminski, A., Krafft, T., Loh, W., Otto, P., & Puntschuh, M. (2020). *From Principles to Practice. An Interdisciplinary Framework to Operationalise AI Ethics*. Bertelsmann Stiftung. https://www.bertelsmann-stiftung.de/de/publikationen/publikation/did/from-principles-to-practice-wie-wir-ki-ethik-messbar-machen-koennen

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., … Wen, J.-R. (2023, Juni 29). *A Survey of Large Language Models*. http://arxiv.org/abs/2303.18223