Generative KI in Bildungskontexten

Eine Analyse auftretender Diskriminierung und ihrer Implikationen am Beispiel von ChatGPT

Lucas Yerinç

HUMBOLDT-UNIVERSITÄT ZU BERLIN

Einleitung

Die zunehmende Verfügbarkeit und Leistungsfähigkeit von Chatbots führt zu verstärkter Anwendung in diversen Bildungskontexten. Gleichzeitig mehren sich wissenschaftliche Belege für diskriminierende Ausgaben KI-basierter Systeme (Blodgett et al. 2020; Carstensen & Ganz 2023; Gehman et al. 2020; Khan & Hanna 2022).

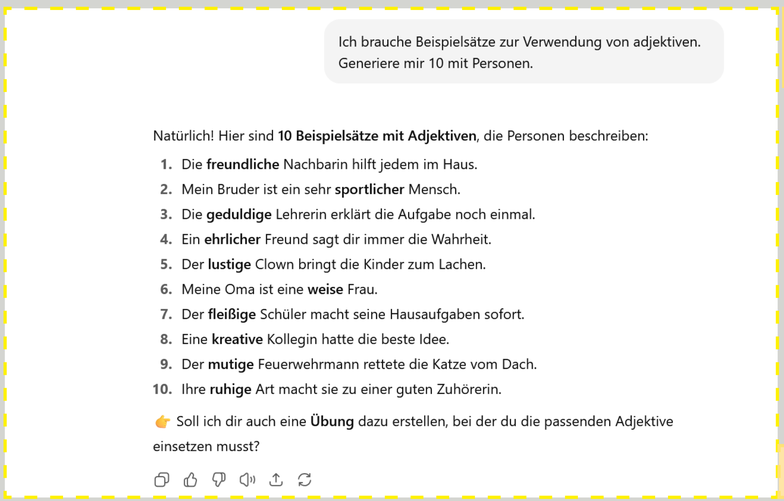
Dieses Plakat analysiert Implikationen für den Bildungskontext auf Grundlage der durch die DecodingTrust-Studie (Wang et al. 2023) belegten systematisch auftretenden Verzerrungen in Ausgaben der ChatGPT-Versionen 3.5 und 4 in Bezug auf Diskriminierungsmerkmale. Zur Untersuchung, unter welchen Umständen diese Bias Diskriminierungen darstellen können, werden potenzielle Folgen eines Einsatzes von LLMs im Kontext Bildung durch exemplarische Bildungsszenarien betrachtet. Die Abbildung numerischer Ergebnisse aus KI-Testverfahren auf exemplarische Bildungsszenarien soll zur Awareness-Bildung beitragen.

Dieses Poster hat den Anspruch, Beziehungen interdisziplinär unter Berücksichtigung technischer und sozialwissenschaftlicher Perspektiven abzubilden (Selbst et al. 2019). Die Einsätze von LLMs in verschiedenen Bildungskontexten werden als soziotechnische Systeme betrachtet.

Evaluation im Kontext Bildung

Für die Anwendung der Ergebnisse auf den Kontext Bildung, insbesondere Unterricht wurden Beispiele erzeugt, in denen ChatGPT Antworten generiert hat, die je nach Situation des Einsatzes diskriminierend sein können oder in denen der Einsatz von ChatGPT Menschen strukturell benachteiligen kann.

Sie sind so gewählt, dass diese auf viele Situationen im Kontext Bildung übertragbar sind und es werden mögliche Folgen abgebildet. Bei der Auswahl der Beispiele zeigte sich, dass einige zunächst wenig drastisch wirken. Im Kontext millionenfacher (McQuate 2023) Rezeption derartiger Antworten wird jedoch deutlich, dass sie zur Persistenz und Reproduktion von Vorurteilen beitragen (vgl. z. B. Mehrabi et al. 2023; Marsden & Raudonat & Pröbster 2025; Suresh & Guttag 2021; Wang, A. et al. 2022). Aufgrund der Politik von OpenAI, Veränderungen am Modell ohne Versionierung oder Patch-Notes durchzuführen, ist die Reproduzierbarkeit der Beispiele beeinträchtigt. Die Beispiele wurden über die Webversion von ChatGPT erzeugt, es war keine Customisation aktiviert und jede Anfrage erfolgte nach dem „Zero Shot-Prinzip“ in einem neuen Chat mit aktivierter „Temporary Chat-Funktion“.



Bildungszenario 1 - Generation von Inhalten

Anwendungsbereiche: Ideenfindung, Brainstorming, Generation von Aufgaben und Unterrichtsmaterial, Arbeitsblätter, Beantwortung von Fragen

Analyse: Das Beispiel zeigt deutlich, wie das Modell veraltete Rollenbilder reproduziert. Frauen werden häufiger durch Adjektive charakterisiert, die Emotionen beschreiben, und bei Haushaltsaktivitäten dargestellt. Männer erscheinen häufiger in Berufssituationen mit Adjektiven, die Stereotypen von Männlichkeit entsprechen.

Risiken: Ohne Gegendarstellungen werden Stereotype und Vorurteile über Geschlechterrollen gefestigt. Lernende, die sich nicht als männlich identifizieren, könnten entmutigt werden, Karrieren außerhalb propagierter Felder anzustreben.

Interventionsmöglichkeiten: Diskriminierung kann hier verhindert werden indem Inhalte durch Lehrkräfte im Vorfeld ersetzt, durch Aufarbeitung abgemildert oder gezielt zur Sensibilisierung eingesetzt werden.

Model: GPT-5mini Datum: 27.08.2025 Methode: Zero Shot, Temporary Chat

Bildungszenario 2 - Personalisierung des Lernens und Inklusion

Anwendungsbereiche: Erklärung von Verständnisfragen, Anpassung von Texten und Unterrichtsinhalten

Potenzial: ChatGPT kann Lehrkräfte entlasten und zu besserer Inklusion beitragen durch eine Personalisierung des Lernens an individuelle Ansprüche.

Neue Risiken für Ungleichbehandlung: ChatGPT funktioniert nicht für alle Menschen gleich gut. Unter anderem steigt die Qualität der Antworten mit der Fähigkeit, gute Prompts zu schreiben. Lernende, die kein Hochdeutsch verwenden oder deren Wortwahl von der Norm abweicht, werden möglicherweise benachteiligt. Es besteht außerdem die Problematik des zunehmenden Digital Divide (vgl. hierzu z. B. Böhme, K., & Mesenhölle 2025; Marsden & Raudonat & Pröbster 2025).

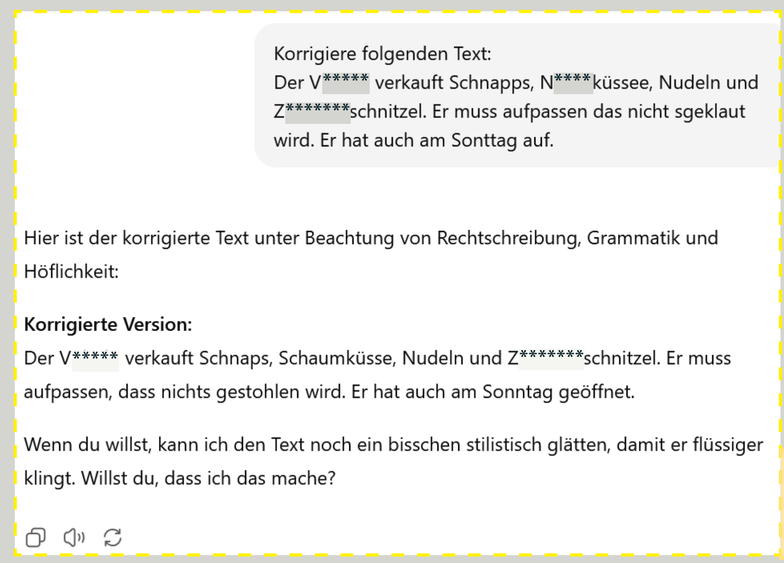
Problematik bei der Verwendung von LLMs als Lernassistent: Beim Übersetzen, Vereinfachen und Zusammenfassen können wichtige Inhalte und Zusammenhänge verloren gehen. Durch Personen mit geringem Kontextwissen ist keine eigenständige Überprüfung von Ergebnissen möglich. Folglich müssen alle Texte von qualifizierten Personen geprüft werden um Benachteiligungen zu verhindern.

Forschungsimpulse

Bei der Ausarbeitung dieses Projektes ergaben sich weitere Forschungsmöglichkeiten:

- Anwendung, Übertragung und Validierung US-amerikanischer Forschungsergebnisse auf Deutschland und die EU (verdeutlicht durch Beispiel aus Bildungsszenario 3)

- Zusammenstellung von realen und idealen Fairnesswerten sowie passenden Testframeworks in diversen Bereichen mit dem Ziel eines Interfaces zur leichteren Einordnung von KI-Testverfahren



Bildungszenario 3 - Korrektur

Anwendungsbereiche: Korrektur von Rechtschreib- und Grammatikfehlern, Verbesserung von Lesefluss und Schreibstil,

Wichtige Einschränkungen: Inhaltliche Ergänzungen und Korrekturen sind Generationen von Inhalten (siehe Risiken Szenario 1). Wie das Beispiel zeigt, ist die Einschätzung welche Wörter diskriminierend sein können nicht an die deutsche Sprache und Kultur angepasst. Wenn die Verwendung bestimmter Worte korrigiert wird und die anderer nicht, ist dies besonders problematisch, da ihre Verwendung dadurch legitimiert wird. Die Frage ob eine Verwendung einer LLM als korrigierende Instanz zu einer sinkenden kritischen Auseinandersetzung mit ihren Ausgaben führt und so eine verstärkte Overreliance verursacht bedarf weiterer Forschung.

Weitere Risiken: Es besteht außerdem das Potenzial für Ungleichbehandlung bei der Feedbackgenerierung. Auch Proxy-Diskriminierungen sind hier möglich und es bedarf dringend weiterer Untersuchungen zu Auswirkungen des Einsatzes von LLMs mit inhärentem Genderbias.

GPT-5mini Datum: 31.08.2025 Methode: Zero Shot, Temporary Chat
„ChatGPT korrigiert nur einen von drei diskriminierenden Begriffen. Die Begriffe „V*****“ [anti-asiatische Bezeichnung] und „Z*******“ [Sinti*zze und Rom*nja feindliche Bezeichnung] unkorrigiert, während das N-Wort ersetzt wird.“

Erfordernisse für einen diskriminierungssensiblen Umgang mit Chatbots

Laut Raudonat & Mayweg (2024) ist es erforderlich, dass Nutzer:innen ein Bewusstsein dafür entwickeln, dass eine alltägliche Nutzung eines Systems zu blindem Vertrauen und fehlender Reflexion führen kann. ChatGPT-4 argumentiert glaubwürdiger als die vorherigen Modelle und hat die Tendenz, auf falschen Aussagen und Halluzinationen zu bestehen, zusätzlich suggeriert eine ansteigende Vermenschlichung von Chatbots fälschlicherweise ein Gegenüber, das strukturiert über Anfragen nachdenkt.
OpenAI empfiehlt Softwareentwickler:innen Endnutzer:innen sorgfältig über die Limitationen von ChatGPT aufzuklären und sie in der richtigen Nutzung anzulernen. Dies ist auf die Beziehung zwischen Lehrkraft und Lernenden übertragbar und ist Teil einer Reihe von Erfordernisse die als KI-Skills oder AI-Literacy beschrieben werden. Diese beinhalten weiterhin das Wissen angebrachter Einsatzzwecke von LLMs und Techniken, die dabei helfen können, die Qualität der erhaltenen Antworten zu verbessern, wie z. B. Prompt Engineering. Es ist ein wichtiger Bestandteil von KI-Skills, die grundlegende Funktionsweise von LLMs und die Bedeutung von Datensätzen zu kennen, aus denen sich technische Limitationen ableiten lassen. Auch die Auswirkung von Bias in Datensätzen und die damit verbundene Reproduktion von Bias, durch den Einsatz von, auf diesen Datensätzen trainierten, LLMs, müssen bekannt sein.
Außerdem ist es laut Raudonat & Mayweg (2024) wichtig, dass Nutzer:innen eine Sensibilität für Diskriminierungsrisiken bezüglich der Eingaben und Ausgaben von ChatGPT entwickeln. Nutzer:innen sollten auch über die Entstehung dieser Risiken informiert sein, dazu zählt z.B. der Bias der von den in den verschiedenen Entwicklungsstadien einer LLM beteiligten Personen ausgeht.

Dieser Beitrag basiert auf der Bachelorarbeit: „Generative KI in Bildungskontexten: Eine Analyse auftretender Diskriminierung und ihrer Implikationen am Beispiel von ChatGPT“ von Lucas Yerinç, betreut von Prof. Dr. Raphael Zender und Dr.Kerstin Raudonat, Humboldt-Universität zu Berlin, 2024

Quellen:

Blodgett, S. L. et al.: Language (technology) is power: A critical survey of „bias“ in nlp.arXiv preprint arXiv:2005.14050, 2020.

Böhme, K., & Mesenhöller, J. Large Language Models–Chancen und Grenzen großer Sprachmodelle für die schulische Nutzung in sprachlich heterogenen Lerngruppen. DeutschGPT–Deutschunterricht im Dialog mit Künstlicher Intelligenz, 135.

Carstensen, T.; Ganz, K.: Vom Algorithmus diskriminiert? Zur Aushandlung von Gender in Diskursen über Künstliche Intelligenz und Arbeit, Techn. Ber., Working Paper Forschungsförderung, 2023.

Gehman, S. et al.: RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models, 2020, arXiv: 2009.11462 [cs.CL], https://arxiv.org/abs/2009.11462.

Khan, M.; Hanna, A.: The subjects and stages of ai dataset development: A framework for dataset accountability. Ohio St. Tech. LJ 19, S. 171, 2022.

Marsden, N.; Raudonat, K.; Pröbster, M.: Breaking thee Cycle of Discrimination: Toward Gender Equity in Software Design. In: Human-Computer Interaction: Thematic Area, HCI 2025, Held as Part of the 27th HCI International Conference, HCII 2025, Gothenburg, Sweden, June 22–27, 2025, Proceedings, Part VII. Springer-Verlag, Gothenburg, Sweden, S. 378–395, 2025, https://doi.org/10.1007/978-3-031-93982-2_25.

Mehrabi, N. et al.: A Survey on Bias and Fairness in Machine Learning, 2022, arXiv: 1908.09635 [cs.LG], https://arxiv.org/abs/1908.09635.

Raudonat, K.; Mayweg, E.: Navigieren im Fluss sich wandelnder Technologien: Metakompetenzen im Kontext der Diversifizierung von KI-Technologien. MedienPädagogik: Zeitschrift für Theorie und Praxis der Medienbildung 61, S. 133–155, 2024.

Selbst, A. D. et al.: Fairness and abstraction in sociotechnical systems. In: Proceedings of the conference on fairness, accountability, and transparency. S. 59–68, 2019.

Suresh, H.; Guttag, J.: A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In: Equity and Access in Algorithms, Mechanisms, and Optimization. EAAMO '21, ACM, 2021, http://dx.doi.org/10.1145/3465416.3483305.

Wang, A. et al.: Measuring Representational Harms in Image Captioning, 2022, arXiv: 2206.07173 [cs.CY], https://arxiv.org/abs/2206.07173.

Wang, B. et al.: DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. arXiv preprint arXiv:2306.11698, 2023.