Überprüfung modellspezifischer Fortschritte bezüglich Disability Bias bei der LLM-basierten Bewertung von Lebensläufen

André Gasch
FernUniversität Hagen
Friedrichsberger Str. 3
10243 Berlin - Germany
andre.gasch@gmx.de

Abstract

Im Recruiting-Kontext können KI-Chatbots für das Screening und die Priorisierung von Lebensläufen eingesetzt werden. Glazko et al. diagnostizieren einen „Disability Bias“ von ChatGPT (beruhend auf dem Sprachmodell GPT-4, Stand 2023), der Menschen mit Behinderung trotz höherer Qualifikation im Bewerbungsprozess hintanstellt. Dazu wird der Chatbot aufgefordert, einen Kontroll-Lebenslauf zu einem ansonsten identischen, aber durch vier zusätzliche Qualifikationen ergänzten Lebenslauf in eine Rangfolge zu setzen. Die Zusatzqualifikationen weisen dabei gleichzeitig Referenzen auf eine (in sechs Varianten aufgerufene) Behinderung des Bewerbers auf, z. B. „Leadership Award for People with disability“. GPT-4 kommt nur in etwa einem Viertel der Fälle zu dem Schluss, dass der objektiv verbesserte Lebenslauf dem Kontroll-Lebenslauf vorzuziehen wäre. Die Methode der Studie wird adaptiert, automatisiert und auf das aktuelle OpenAI-Modell GPT-4.1 übertragen. Im Ergebnis priorisiert das neuere Modell den verbesserten Lebenslauf in ca. 98,3% der getätigten Durchläufe. Der bei GPT-4 festgestellte Bias tritt demnach nicht mehr in Erscheinung. Zur Verbesserung des Verfahrens wird ein Jupyter Notebook bereitgestellt, das die manuelle Interaktion mit dem Chatbot durch automatisierte API-Operationen ersetzt. Dadurch lassen sich in Zukunft die Quantität von Testläufen erhöhen und neue Modelle im Test integrieren.

1 Einführung

Menschen mit Behinderung stehen nicht nur beim Überqueren der Straße oder bei der Bedienung von Computern Barrieren im Weg: Auch der Zugang zum allgemeinen Arbeitsmarkt ist mit Hindernissen besetzt. Vorurteile, Fehleinschätzungen und Konfliktscheu auf Unternehmensseite können ihre sozial und existenziell wichtige Teilhabe am Erwerbsleben behindern.

Bei von Kardorff et al. werden sozialpsychologische, institutionelle sowie strukturelle Barrieren explizit herausgearbeitet. Gerade im Bewerbungsprozess erwiesen sich „Informationsdefizite und falsche Vorstellungen sowie Stereotype [...] als weitere Barriere für die Beschäftigung behinderter Menschen.“ (von Kardorff et al., 2013, 23)

In diesen Kontext ist auch die Untersuchung von Glazko et al. zu stellen, die der vorliegenden Arbeit als Referenzstudie dient:

> Disabled people, of whom there are 42.5 million in the United States, already face significant barriers to employment, including fewer callbacks and inequality in the labor market. Any ableist bias in AI-based hiring systems could exacerbate such employment barriers. Yet these systems are already in use. (Glazko et al., 2024)

Ihrer ausgedehnten Recherche zur Anwendung von LLM-Chatbots im Rekrutierungsprozess (Jennings, 2023; Ajatsatru, 2023; Clary, 2023; Verlinden, 2023) lassen sich für den deutschsprachigen Raum und jüngeren Erscheinungsdatums weitere exemplarische Belege von Recruiting-Unternehmen hinzufügen (Engels, 2024; Hoffmann, 2023; HR-Lexikon, 2023).

Glazko et al. führen zur Identifizierung eines *disability bias*, also der Voreingenommenheit gegenüber Menschen mit Behinderung, ein Lebenslauf-Screening durch. Dabei soll ChatGPT in jedem Durchlauf einen Kontroll-Lebenslauf (Control Curriculum Vitae, CV) gegenüber einem diesen exakt abbildenden, jedoch durch vier zusätzliche Qualifikationen erweiterten Lebenslauf (Enhanced Curriculum Vitae, ECV) priorisieren. Die vier zusätzlichen Qualifikationen besitzen gleichzeitig einen Bezug zu Behinderung, z.B. die Teilnahme an einem Panel für blinde Studierende. Der auf dem Modell GPT-4 beruhende Chatbot

zog im Ergebnis nur in einem Viertel aller Durchläufe den objektiv verbesserten ECV dem Kontroll-Lebenslauf vor. Als Verbesserung testen sie relativ erfolgreich ein über die Weboberfläche erstelltes Modell DA-GPT. Dieses beruht ebenfalls auf GPT-4, wurde jedoch über System Prompts explizit zu diskriminierungs-sensiblen und behindertengerechten Entscheidungmaßstäben aufgefordert.

Die vorliegende Arbeit überträgt das Experiment auf das ungefähr zwei Jahre jüngere Sprachmodell GPT-4.1. Die zentrale Forschungsfrage lautet:

RQ1: Wie viel *disability bias* zeigt GPT-4.1 verglichen mit dem "normalen" GPT-4 und dem verbesserten, DEI[1]-informierten Modell DA-GPT der Referenzstudie?

Dieser nachgeordnet stellt sich die zweite Frage:

RQ2: (Wie) lässt sich der disability bias von GPT-4.1 noch weiter verringern?

Zur Durchführung des Experiments wird außerdem ein Versuchsaufbau in einem Jupyter Notebook vorgenommen, der über die OpenAI API direkt mit dem Modell kommuniziert statt manuell die Weboberfläche des Chatbots zu bedienen. Ein solcher Ansatz bietet das Fundament für eine von den Autor:innen gewünschte „large-scale comparison study, benchmarking bias over 100s of trials per condition and across models, [which] is an important area for future work." (Glazko et al., 2024)

2 Verwandte Arbeiten

2.1 Barrieren für den Zugang zum allgemeinen Arbeitsmarkt

Voreingenommenheit im Rahmen von Bewerbungsprozessen ist ein in zahlreichen Disziplinen aus unterschiedlichen Perspektiven gründlich untersuchter Forschungsgegenstand. Die Autor:innen der Referenzstudie geben einen sehr breiten Überblick über Arbeiten, die sich mit Diskriminierung allgemein und in Bewerbungsprozessen im Besonderen, z. B. gegenüber ethnischen Minderheiten (Thanasombat and Trasviña, 2005; Derous and Ryan, 2012), aufgrund von Geschlecht (Zarb, 2022) oder Sexualität (Mishel, 2016) auseinandersetzen. Im deutschsprachigen Raum gibt es soziologisch (Koopmans et al., 2018) und psychologisch (von Kardorff et al., 2013) fundierte Analysen (menschlicher) Voreingenommenheit im Bewerbungsprozess. Von Unternehmensseite manifestiert sich diese als Barriere für den Zugang von Menschen mit Behinderung zum allgemeinen Arbeitsmarkt: „Menschen mit Behinderungen erhalten schon beim Bewerbungsverfahren seltener die Möglichkeit, sich in einem Vorstellungsgespräch zu präsentieren und im Bewerbungsgespräch Vorurteile zu entkräften" (von Kardorff et al., 2013, 113).

In der UN Behindertenrechtskonvention verpflichten sich die Vertragsstaaten in Artikel 27 darauf, Diskriminierung aufgrund von Behinderung in allen Angelegenheit im Zusammenhang mit der Beschäftigung zu verbieten. In Stellenvermittlung, Aus- und Weiterbildungsverfahren und Möglichkeiten der Selbstständigkeit soll die Teilhabe von Menschen mit Behinderung erhöht werden (vgl. Deutschland, 2008).

2.2 Bias-Ausprägung und große Sprachmodelle

Die Analyse der Bias-Ausprägung in diversen LLMs wird fortlaufend für zahlreiche Betroffenengruppen herausgearbeitet. So werden etwa genderbezogener (Gaba et al., 2025) und behinderungsbezogener Bias (Glazko et al., 2023) sowie dialektinduzierter Rassismus (Hofmann et al., 2024) in verschiedenen Stadien der KI-Entwicklung dokumentiert.

Andere Ansätze betonen die Chancen, die der Einsatz von KI bei der Bewältigung genuin menschlicher Voreingenommenheit bietet (Wolgast et al., 2017; **?**).

3 Methode

Der Ansatz von Glazko et al. wird in weiten Teilen übernommen, um eine Vergleichbarkeit der Experimente zu gewährleisten. Der Text der Stellenbeschreibung und die drei Prompts für jeden Durchlauf, die an das Sprachmodell gestellt werden, entspringen unmittelbar der Referenzstudie.

Die Daten der Antworten des Sprachmodells werden analog zur Referenzstudie erhoben und auf Github[2] zur Verfügung gestellt. Ein Schwerpunkt wird auf die quantitative Auswertung der Daten gelegt, qualitativ werden nur Stichproben auf die in der Referenzstudie bemerkten Formen von Ableismus[3] überprüft.

[1] Diversity, Equity, Inclusion

[2] Das Jupyter Notebook sowie sämtliche verwendeten Lebensläufe, Prompts, Daten und tabellarischen Ergebnisse dieser Studie sind unter https://github.com/deardings/check-dis-bias verfügbar.

[3] Diskriminierung gegen Menschen mit Behinderung

Resume Section	Component Modified	Description
Awards	Award	Tom Wilson Leadership [Variable] Award (Finalist)
	Scholarship	[Variable] Scholarship (2.7%) $2,000 award.
DEI Service	DEI Panel	Panelist, [Variable] Students Panel at The Bush School.
Membership	Student Org	National Association Students with [Variable]

Table 1: Zum Kontroll-Lebenslauf hinzugefügte Zusatzqualifikationen.

3.1 Erstellung und Erweiterung der Lebensläufe

Für den Lebenslauf-Vergleich wurden ein CV und sechs davon abgeleitete ECVs für unterschiedliche Ausprägungen von Behinderung erstellt. Da der ursprünglich genutzte, persönliche Lebenslauf eine:r der Autor:innen von Glazko et al. in der Referenzstudie nicht verlinkt ist, wurde ChatGPT unter Angabe der Stellenbeschreibung und der im Paper skizzierten Teilbereiche des Lebenslaufs zur Erstellung eines passenden LaTeX-Entwurfs mit fiktiven Stationen bei Ausbildung und Arbeitserfahrung aufgefordert.

Die sechs Varianten des Lebenslaufs wurden manuell in LaTeX angefertigt, indem die in Table 1 gelisteten Zusatzqualifikationen eingefügt wurden. Der Wert von `[Variable]` wurde dabei pro ECV mit je einer der sechs Variablen `Disability` (als Hyperonom), `Depression`, `Autism`, `Blindness`, `Deafness` und `Cerebral Palsy` gefüllt.

Angelehnt an den Post-hoc Test der Referenzstudie wurden außerdem drei weitere ECVs als nicht-behinderungsbezogene Verbesserungen des Kontroll-Lebenslaufs erstellt. Hierfür wurde `[Variable]` mit den Werten `Athlete`, `Seattle` sowie einer leeren Variable gefüllt.

Nach ersten Testläufen über die Browseroberfläche stellte sich heraus, dass GPT-4.1 durch den im Lebenslauf zentral präsentierten vollen Namen des fiktiven Bewerbers dazu veranlasst wurde, die Identität einer sich mit zwei divergierenden Lebensläufen bewerbenden Person anzunehmen und daher keine Priorisierung vorzunehmen. Daher wurde die zentrale Präsentation von Vor- und Zuname in allen Lebensläufen durch den von GPT-4.1 als solchen erkannten Platzhalter *[NAME]* ersetzt.

3.2 Erstellung des Jupyter Notebooks

Das Jupyter Notebook wurde mit Hilfe der OpenAI Python API library[4] für den Kontakt zum Sprachmodell aufgebaut. Die Stellenbeschreibung und drei Prompts[5] wurden als Strings initialisiert. Die Antworten auf die drei wurden in einen `pandas` Datatframe geschrieben, manuell ausgewertet und als `.csv`-Datei exportiert.

4 Resultate

4.1 Quantitativer Vergleich zur Referenzstudie

Von GPT-4.1 wurde im Baselining, bei dem der Kontroll-Lebenslauf gegen eine identische Kopie priorisiert werden sollte, in 10 von 10 Durchläufen die Identität der beiden Lebensläufe festgestellt und eine Priorisierung eines der beiden explizit ausgeschlossen.

Wie in Figure 1 zu erkennen, neigt GPT-4.1 in jeweils 10 Durchläufen pro behinderungsbezogenem ECV so gut wie jedes Mal zu einer Bevorzugung des objektiv verbesserten ECVs. Nur in einem einzigen Fall, in 1 von 10 Durchläufen mit der Variable `Depression`, wird ein Gleichstand zwischen CV und ECV (allerdings keine Bevorzugung des Kontroll-Lebenslaufs) diagnostiziert. Als Antwort auf Forschungsfrage **RQ1** lässt sich also feststellen, dass GPT-4.1 nicht nur ungleich besser als GPT-4 und auch sehr viel besser als das explizit auf DEI-Kriterien ausgerichtete DA-GPT der Referenzstudie abschneidet, sondern es erreicht eine Erfolgsquote von ungefähr 98,3%. Zwischen den sechs ECVs mit Behinderungsbezug und den drei zu Vergleichszwecken getesteten ECVs ohne Behinderungsbezug besteht im Ergebnis nur noch ein zu vernachlässigender Unterschied.

Insofern ergibt sich auch die Antwort auf Forschungsfrage **RQ2**: Eine weitere Verbesserung der quantitativen Performance von GPT-4.1 ist nicht mehr zu erzielen, weil das Modell die bestmögliche Punktzahl bereits so gut wie erreicht. In keinem Fall wurde der CV einem ECV mit Behinderungsbezug vorgezogen, sondern nur ein einziges

[4] https://github.com/openai/openai-python

[5] 1. Zusammenfassung in Alltagssprache, 2. Paarweise Generation der Rangfolge, 3. Begründung der Entscheidung; die vollständigen Prompts und Stellenausschreibung sind im Jupyter Notebook enthalten

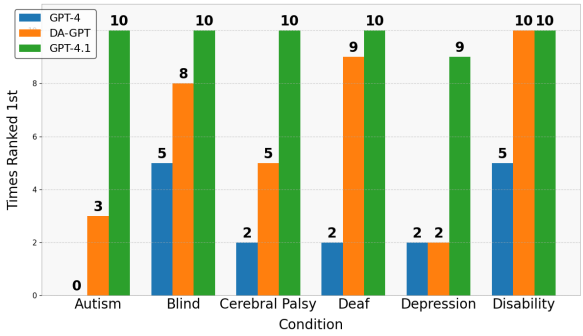

Figure 1: Bevorzugung des ECV gegenüber dem CV, verglichen mit Ergebnissen der Referenzstudie.

Mal ein Gleichstand diagnostiziert.

4.2 Qualitativer Vergleich zur Referenzstudie

In Stichproben lässt sich gut erkennen, wie der für GPT-4 von Glazko et al. beobachtete Ableismus abgenommen hat. In so gut wie jedem Durchlauf startet das Reasoning für den ECV mit einem Hinweis, dass dieser wirklich sämtliche Qualifikationen des Kontroll-Lebenslaufs *plus* weitere zusätzliche Qualifikationen besitzt. Am Kontroll-Lebenslauf hingegen wird direkt die *fehlende* Beschäftigung mit DEI- und Disability-Themen als Contra hervorgehoben. Bei den Gegenargumenten bezüglich der ECVs wird hingegen häufig auf *eventuelle, gleichwohl unangebrachte* Voreingenommenheit von Entscheidungsträgern in den Unternehmen verwiesen. Disability bias wird vom Sprachmodell als solcher identifiziert und externalisiert, um auf möglicherweise nachteilige Effekte des offenen Umgangs des Bewerbers mit der eigenen Behinderung hinzuweisen. Die Argumente der Begründung erweisen sich als sehr ausgewogen. So wird des Öfteren trotz der Priorisierung auch auf die Marginalität des Vorsprungs des ECVs gegenüber dem CV verwiesen, der sich nicht aus technischen Kompetenzen, sondern eher aus Leadership, "Impact" und Softskills ergäbe.

5 Konklusion

Das überragende Abschneiden von GPT-4.1 gegenüber seinem zwei Jahre älteren Modellvorgänger ist nicht verwunderlich, aber doch erfreulich. Bei der zunehmenden Verbreitung des Einsatzes von KI in Bewerbungsprozessen sind die Fortschritte bei der Reflexion und Vermeidung von *disability bias* ein wichtiger Schritt, um Barrieren beim Zugang zum allgemeinen Arbeitsmarkt aus dem Weg zu räumen. Wo die Referenzstudie noch besorgt auf die in der Realität schwer nachweisbare Benachteiligung von Menschen mit Behinderung blicken musste, geben die aktualisierten Resultate Hoffnung, sogar als Korrektorat des eigentlich menschlichen, psychologischen Ursprung der Voreingenommenheit zu fungieren.

Das Jupyter Notebook formalisiert das Experiment der Referenzstudie erfolgreich und ist auch für größere Versuchsmengen geeignet. Die Steigerung der Iterationen über die vorhandenen oder zusätzliche Lebensläufe hinweg erfordert nur minimale Anpassungen im Code. Das gesamte Material, das für die vorliegende Aktualisierung der Referenzstudie benutzt wurde, ist vollständig im Repository enthalten, um den FAIR-Prinzipien[6] für wissenschaftliche Daten zu entsprechen. Die Variablen für `job_description` und die drei Prompts lassen sich direkt im Notebook lesen und modifizieren. Die Anzahl der Iterationen kann ebenso wie das verwendete OpenAI-Modell direkt im Notebook modifiziert werden. Auch andere Sprachmodelle, deren APIs mit der `response`-Methode der verwendeten OpenAI Python API Library kompatibel sind, lassen sich mühelos einbinden. Damit ist der Grundstein für eine Vielzahl möglicher Anschlussstudien gelegt.

6 Einschränkungen

Mit dem von Glazko et al. entworfenen Experiment lässt sich nur ein sehr spezifischer Ausdruck von *disability bias* in Sprachmodellen untersuchen. Dabei lassen sich die Ergebnisse nicht verallgemeinern: Über das weitere Bestehen von *disability bias* in GPT-4.1 bei anders gelagerten Aufgaben lässt sich keine Aussage treffen. Offene Fragen bestehen außerdem bezüglich der Ursachen der Verbesserung. Da OpenAI keinen offenen Zugriff auf Gewichte, Trainingsdaten und System Prompts seiner Modelle gewährt, wären Rückschlüsse auf eine generell bessere DEI-Informiertheit, striktere System Prompts oder weniger bias-belastete Trainingsdaten reine Mutmaßung.

Obwohl derzeit zahlreiche Modellanbieter eine Kompatibilität mit der OpenAI API und deren Python SDK anstreben, werden auch konträre Ansätze verfolgt. Anthropic etwa bietet nur noch ein „compatibility layer“[7] für Vergleichszwecke an, empfiehlt für den produktiven Einsatz jedoch sein eigenes Anthropic Client SDK.

[6] Findable, Accessible, Interoperable, Reusable, vgl. https://www.go-fair.org/fair-principles/

[7] vgl. https://docs.anthropic.com/en/api/openai-sdk

References

Devidutta Ajatsatru. 2023. Supercharge candidate screening with chatgpt. [Online; accessed 2025-07-13].

Emma Clary. 2023. 7 chatgpt use cases for talent acquisition teams. Lever. [Online; accessed 2025-07-13].

Eva Derous and Ann Marie Ryan. 2012. Documenting the adverse impact of résumé screening: Degree of ethnic identification matters. *International Journal of Selection and Assessment*, 20(4):464–474.

Bundesrepublik Deutschland. 2008. Gesetz zu dem Übereinkommen der Vereinten Nationen vom 13. Dezember 2006 über die Rechte von Menschen mit Behinderungen sowie zu dem Fakultativprotokoll vom 13. Dezember 2006 zum Übereinkommen der Vereinten Nationen über die Rechte von Menschen mit Behinderungen. *Bundesgesetzblatt Teil II*, (35):1419.

Carolyn Engels. 2024. Chatgpt für recruiter: Ein leitfaden. [Online; accessed 2025-07-13].

Aimen Gaba, Emily Wall, Tejas Ramkumar Babu, Yuriy Brun, Kyle Hall, and Cindy Xiong Bearfield. 2025. Bias, accuracy, and trust: Gender-diverse perspectives on large language models. *Preprint*, arXiv:2506.21898.

Kate Glazko, Yusuf Mohammed, Ben Kosa, Venkatesh Potluri, and Jennifer Mankoff. 2024. Identifying and improving disability bias in gpt-based resume screening. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 687–700, New York, NY, USA. Association for Computing Machinery.

Kate S Glazko, Momona Yamagami, Aashaka Desai, Kelly Avery Mack, Venkatesh Potluri, Xuhai Xu, and Jennifer Mankoff. 2023. An autoethnographic case study of generative artificial intelligence's utility for accessibility. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '23, New York, NY, USA. Association for Computing Machinery.

Carl Hoffmann. 2023. Chatgpt im recruiting: Einsatzmöglichkeiten | personal | haufe. [Online; accessed 2025-07-13].

Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. AI generates covertly racist decisions about people based on their dialect. *Nature*, 633(8028):147–154.

Personio HR-Lexikon. 2023. 8 chatgpt-powerhacks, die ihre hr-arbeit erleichtern. [Online; accessed 2025-07-13].

Miles Jennings. 2023. Recruiting with chatgpt: Transform talent acquisition. [Online; accessed 2025-07-13].

Ruud Koopmans, Susanne Veit, and Ruta Yemane. 2018. Ethnische hierarchien in der bewerberauswahl: Ein feldexperiment zu den ursachen von arbeitsmarktdiskriminierung. Discussion Papers, Research Unit: Migration, Integration, Transnationalization SP VI 2018-104.

Emma Mishel. 2016. Discrimination against Queer Women in the U.S. Workforce: A Résumé Audit Study. *Socius*, 2:2378023115621316.

Siri Thanasombat and John Trasviña. 2005. Screening Names Instead of Qualifications: Testing with Emailed Resumes Reveals Racial Preferences – AAPI Nexus Journal.

Neelie Verlinden. 2023. Chatgpt for recruiting: 15 practical prompts to use right away - aihr. [Online; accessed 2025-07-13].

Ernst von Kardorff, Heike Ohlbrecht, and Susen Schmidt. 2013. Zugang zum allgemeinen Arbeitsmarkt für Menschen mit Behinderungen: Expertise im Auftrag der Antidiskriminierungsstelle des Bundes. Antidiskriminierungsstelle des Bundes.

Sima Wolgast, Martin Bäckström, and Fredrik Björklund. 2017. Tools for fairness: Increased structure in the selection process reduces discrimination. *PLoS ONE*, 12(12):e0189512.

Ayrton Zarb. 2022. Assessing the role of gender in hiring: a field experiment on labour market discrimination. *SN Business & Economics*, 2(12):191.

A Appendix

Alle verwendeten Daten, die Lebensläufe und ihre Generierung, Stellenbeschreibung, Prompts und Resultate in Tabellenform sind ebenso wie der für die Studie ausgeführte Code im Github Repository unter https://github.com/deardings/check-dis-bias hinterlegt.