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

Abstract

Glazko et al.[1] haben eine Studie über die GPT-basierte Bewertung von Lebensläufen vorgelegt, in der mit einer eingängigen Methode ein „Disability Bias“ von ChatGPT (beruhend auf dem Modell GPT-4) diagnostiziert wird. Sie quantifizieren diesen Bias in Bezug auf unterschiedliche Formen von Behinderung und stellen ihm die deutlich besseren Ergebnisse eines explizit DEI(Diversity, Equity and Inclusion)-informierten GPT-4-Modells gegenüber. Angesichts der rasanten Weiterentwicklung verschiedener Sprachmodelle sind Fortschritte beim Abbau solcher Ungleichgewichte und bei den Maßnahmen zu ihrer Korrektur immer wieder neu auf den Prüfstand zu stellen. Die Methode der Studie wird daher adaptiert, um sie auf das aktuelle OpenAI-Modell GPT-4o sowie das Reasoning-Modell GPT-o1 anzuwenden. Die Ergebnisse legen einen deutlichen Rückgang des Bias und eine implizite Sensibilität der Modelle gegenüber Behinderungen nahe, die an das explizit DEI-informierte Modell von Glazko et al. heranreicht. Als wesentliche Verbesserung der Verfahrensweise wird ein Jupyter Notebook entwickelt, das die manuelle Interaktion mit dem Chatbot durch automatisierte API-Operationen ersetzt. Dadurch lässt sich in Zukunft die Quantität von Testläufen erhöhen. Zudem können neue Modelle mit geringer Anpassung unter reproduzierbaren Bedingungen untersucht werden.

[1] Kate Glazko, Yusuf Mohammed, Ben Kosa, Venkatesh Potluri, and Jennifer Mankoff. 2024. Identifying and Improving Disability Bias in GPT-Based Resume Screening. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’24), June 03–06, 2024, Rio de Janeiro, Brazil*. ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3630106.3658933.