Erkennung menschlicher und maschineller Autorschaft – ein DEDICAITE-Projekt

Steffen Hessler & Maria Berger, Ruhr-Universität Bochum

Die zunehmende Nutzung generativer KI durch Studierende stellt Hochschullehrende vor neue Herausforderungen. Frühere Studien zeigten, dass Fachleute aus Medizin und Geisteswissenschaften zu 70 % erkennen konnten, ob ein Text von Studierenden oder von ChatGPT stammt. In einer randomisierten Studie überprüften wir diese Trefferquote mit 295 Lehrenden aus allen Fakultäten und untersuchten, ob linguistische Merkmale wichtiger sind als inhaltliche.

Dazu erhielten die Teilnehmenden je einen studentisch oder von ChatGPT-4.0 verfassten akademischen Text. Eine Gruppe bekam Hinweise zu sprachlichen Erkennungsmerkmalen, die andere nicht. Die Erkennungsraten lagen bei 66 % bzw. 63,8 % (nicht signifikant), obwohl nur 11 % ein vertrautes Thema erhielten. Bei Nicht-Geisteswissenschaftler*innen führten die Hinweise zu einer deutlich höheren Trefferquote (75 % vs. 59 %). Insgesamt wurden menschlich verfasste Texte häufiger korrekt erkannt (72 % vs. 58 %).

Um die sprachlichen Unterschiede systematisch zu erfassen, setzte die Forschungsgruppe Verfahren aus der forensischen Linguistik und der Autorschaftserkennung ein. Dabei wurden sprachliche Merkmale analysiert, die typischerweise in menschlichen beziehungsweise KI-generierten Texten vorkommen. So konnten Indikatoren identifiziert werden, die mit hoher Wahrscheinlichkeit auf menschliche oder maschinelle Autorschaft hinweisen, sowie Merkmale, die sich als weniger zuverlässig erwiesen. Diese Erkenntnisse tragen dazu bei, die Erkennung von KI-generierten Texten weiter zu verfeinern und ein besseres Verständnis für die sprachlichen Unterschiede zwischen menschlichen und KI-verfassten Texten zu entwickeln.