Positionspapier: Nutzung Künstlicher Intelligenz in der Evaluation

Diskussionsvorlage zur Session „Künstliche Intelligenz in der Evaluation: Wie positioniert sich die Fachcommunity?“ des AK Professionalisierung, DeGEval Jahrestagung 2025 (Stand: 15.9.2025)

Abstract: Künstliche Intelligenz (KI) hat das Potenzial, die Evaluationspraxis in vielerlei Hinsicht zu verändern, wie aktuell im Falle der Large Language Models (LLM) und der auf ihnen basierenden KI-Chatbots sichtbar wird. Eine pauschale Bewertung des KI-Einsatzes fällt aufgrund der dynamischen Entwicklung, bestimmten Eigenarten der Technologie sowie der Breite möglicher Anwendungen in Evaluationen schwer. Dennoch kann die Frage, welche Formen des Einsatzes in welchem Maße als zulässig gelten, nicht alleine einer emergenten Praxis und dem Spiel der Marktkräfte überlassen werden. Im Bewusstsein, dass die Beantwortung vieler Fragen noch mehr empirische Erfahrungen und fachinterne Diskussionen erfordern und daher nur vorläufig zu entscheiden sind, soll diese Vorlage eine erste Orientierung geben und Fixpunkte für die weitere Entwicklung festhalten.

Inhalt

1 Der KI-Einsatz sollte aus Evaluationsperspektive bewertet werden.

- KI überrascht trotz immer noch auftretender Fehlleistungen viele mit frappierend überzeugenden und augenscheinlich guten Ergebnissen. Daraus entspringen oft ***hohe Erwartungen*** hinsichtlich möglicher positiver Effekte beim Einsatz für Evaluationen.
- Entscheidungen über den KI-Einsatz setzen aber eine ***wohlabgewogene Bewertung*** voraus.
- ***Professionelle Evaluation*** darf diese Bewertungen nicht auf Basis von Augenscheinevidenzen, unrepräsentativen Positivbeispielen oder selektiven Kriterien fällen. Die Entscheidung sollte vielmehr Ergebnis eines Bewertungsprozesses sein, der umso systematischer sein muss, je zentraler die Rolle von KI sein soll.
- Die folgenden Punkte versuchen diesem Anspruch Rechenschaft zu tragen.

2 Zentralität und Art des KI-Einsatzes sind elementar für die Bewertung.

- KI kann potenziell für äußerst ***unterschiedliche Zwecke*** in Evaluationen eingesetzt werden. Sie reichen von einfacher redaktioneller Assistenz, über Unterstützung bei Recherchen, theoretischen Überlegungen oder Schreibprozessen bis hin zu unterschiedlichen Rollen bei Datenerhebung, Datenauswertung und Berichtlegung reichen.
- Bei jeder dieser Einsatzformen sind ***drei Aspekte*** zentral für die Frage, ob der Einsatz zu rechtfertigen ist oder wie kritisch er zu sehen ist:
    - Wie ***zentral*** sind sie KI-unterstützten oder –generierten Inhalte für die Evaluationsqualität? Hier reicht das Spektrum von relativ unkritischer redaktioneller Assistenz bis hin zur Datenauswertung oder Texterstellung durch KI, für die selbst mit starker Ergebniskontrolle (s.u.) die Bedingungen in derzeit noch ungeklärtem Maße gegeben sind.
    - In welchem Maße werden KI-Ergebnisse vor ihrer Verwendung mit menschlicher Evaluationsexpertise ***überprüft, kontrolliert oder im Wechselspiel*** generiert?
    - Erfolgt der Einsatz unter Berücksichtigung ***existierender Regularien*** wie z.B. der Einhaltung von Datenschutz oder Vertraulichkeit?

3 Die Ergebnisqualität ist ein wichtiges Kriterium, darf aber nicht das einzige sein.

- Bei der Frage der ***Zulässigkeit und Sinnhaftigkeit*** des KI-Einsatzes steht oft die Frage im Vordergrund, ob KI-generierte Ergebnisse, z.B. bei Datenauswertungen, menschlich generierten gleichwertig oder gar überlegen sind.
- Entscheidungen über den Einsatz aber alleine davon und von einer erwarteten Aufwandsreduktion abhängig zu machen, würde aber alleine Outputs berücksichtigen. Eine ***umfassende Bewertung*** sollte sich aber wie jede Evaluation nie alleine auf Outputs stützen.
- Weitere Gesichtspunkte sind etwa:
    - ***Inputs***: z.B. Bedingungen der KI-Nutzung wie urheberrechtliche Verstöße der meisten Modellanbietenden
    - ***Prozesse***: z.B. Black-box-Charakter der Output-Generierung oder Schwierigkeiten bei der transparenten Dokumentation von Methoden
    - ***Outcomes***: z.B. mögliche Auswirkung der Verwendung von KI-generierten Ergebnissen etwa auf Glaubwürdigkeit der Evaluation
    - ***Impacts***, z.B. weiterführende Auswirkungen auf den Evaluationsmarkt oder das Beziehungsgefüge in Evaluationen
- So sollte bei den häufig erwarteten Kosten- bzw. Aufwandsersparnissen (Effizienzargument) bedacht werden, ob eingesparte Mittel letztlich der ***Evaluationsqualität*** zugute kommen oder schlicht zur ***Schrumpfung von Evaluationsbudgets*** führen, die ohnehin bereits in der Regel knapp bemessen sind.

4 Der Vergleich mit menschlich generierten Ergebnissen taugt nur bedingt.

- In Beispielen und Studien lässt sich teils bereits eine ***Gleichwertigkeit oder gar Überlegenheit*** der Ergebnisqualität von KI gegenüber menschlichen Aufgabenbearbeitungen belegen. Dies sollte aber aus mehreren Gründen nicht zu voreiligen Schlüssen führen.
- Wie bei anderen Evaluationsergebnissen auch darf es nicht zu einer ***unüberlegten Generalisierung*** von Befunden kommen. KI-generierte Ergebnisse sind von einer großen Zahl möglicher Einflussfaktoren (u.a. Themenfeld, Aufgabenart, Auftragsformulierung, KI-Modell, KI-Anwendung) abhängig, daher ist der lineare Schluss von Erfolgsfällen auf zukünftige Anwendungen nicht voraussetzungslos zulässig.
- KI-Modellen fehlt ein ***Problemverständnis*** im menschlich-kognitiven Sinne. Trotz aller Verfeinerungen arbeiten LLM immer noch stochastisch auf Basis der im Modelltraining verarbeiteten Texte.
- Das Verhalten von KI-Anwendungen wie Chatbots, aber auch dezidierter Software für Spezialaufgaben wie die Datenauswertung, wird in starkem Maße von Bedingungen beeinflusst, ***die nicht offen liegen*** (z.B. Art und Aufbereitung von Trainingsdaten, System Prompts, Orchestrator, Router).
- Zwar sind auch Menschen von nicht dokumentierten Bedingungen in ihrer Aufgabenbearbeitung beeinflusst (z.B. Vorwissen, Vorurteile). Für diese Problematik wurden aber in der Vergangenheit forschungs- und evaluationsmethodologische Lösungen gefunden, die das ***erforderliche Vertrauen*** rechtfertigen. Entsprechende Lösungen können derzeit für den KI-Einsatz noch gar nicht bestehen.
- Ein kritischer grundsätzlicher Faktor ist dabei, dass Menschen grundsätzlich zu ihren Aufgabenbearbeitungen ***Auskunft geben*** und für Fehlleistungen ***zur Rechenschaft gezogen*** werden können. KI-Modelle können beides nicht, weil sie bisher über keine Metakognition verfügen und die Offenlegung bewertungsrelevanter Faktoren von den Softwareanbietenden in der Regel als Betriebsgeheimnis verweigert werden.

5 Ein KI-Einsatz sollte transparent und nachvollziehbar dokumentiert werden.

- Soweit ein KI-Einsatz erfolgt, sollte dieser um gängigen Evaluationsstandards zu entsprechen transparent gemacht werden und so umfassend dokumentiert werden, dass das Zustandekommen von Ergebnissen immer nachvollziehbar ist.
- Die Dokumentation sollte je nach Zentralität des Einsatzes (s.o.) angepasst werden. Als Beispiel für entsprechend gestafelte Dokumentationsrichtlinien können die Hinweise *Zeitschrift für Evaluation* zum KI-Einsatz bei der Manuskriptgestaltung dienen (https://www.degeval.org/zeitschrift-fuer-evaluation/hinweise-fuer-autor-innen/).

6 Evaluationsbeteiligte sollten Anpassungs- und Regulierungsbedarfe prüfen.

KI kann auf verschiedenen Ebenen Nachsteuerungsbedarfe auslösen, die v.a. institutionell geprüft werden sollten.

- ***Auftraggebende*** von Evaluationen sollten prüfen, in welchem Ausmaß und unter welchen Bedingungen sie *die Nutzung von KI* in einer Evaluation akzeptieren. Zusätzlich sollten sie steuern, inwiefern *publizierte Evaluationsergebnisse* für das Training von KI-Modellen genutzt werden können und dürfen. Wenn Auftraggebende KI im *Ausschreibungsprozess* nutzen, sollte dies in angemessener Form transparent gemacht werden, v.a. wenn davon die Auswahl von Angeboten betroffen ist.
- ***Evaluierende*** benötigen relevante Kompetenzen und Ressourcen, um eine fachlich angemessene Entscheidung über den KI-Einsatz zu treffen und diesen kompetent durchzuführen. Dies umfasst auch die Frage der Dokumentation des Einsatzes nach wissenschaftlichen Standards.
- ***Publikationsorgane*** für den wissenschaftlichen und fachlichen Austausch über Evaluation benötigen soweit noch nicht erfolgt ebenfalls eine fachliche Position zur akzeptablen KI-Nutzung bei der Erstellung und Begutachtung von Manuskripten.
- ***Institutionen der Aus- und Weiterbildung*** von Evaluierenden sollten ihr Angebotsportfolio so erweitern, dass Evaluierende als Mindestziel die Fähigkeit erwerben können, gegenstandsangemessen über den Einsatz von KI in Evaluationen zu entscheiden und diesen bewerten zu können. Dies gilt analog für die organisationsinterne Weiterbildung.

7 Wichtig ist nicht die Kostenersparnis, sondern die Qualität von Evaluation.

Unter wettbewerblichen Bedingungen des Evaluationsmarkts können antizipierte Kostenersparnisse eines KI-Einsatzes eine ***problematische Dynamik*** entfalten. Evaluationsanbietende könnten einerseits ihre Angebote günstiger kalkulieren, wenn sie für arbeitsintensive Tätigkeiten den KI-Einsatz planen. Andererseits könnten Auftraggebende versucht sein, entsprechende Angebote aufgrund des günstigeren Preises zu bevorzugen oder gar im Rahmen des Ausschreibungsverfahrens einzufordern.

Die Entscheidung über Art und Ausmaß der KI-Nutzung darf aber nie von der Kostenseite her getrieben sein, sondern muss sich immer am ***Anspruch an die Evaluationsqualität*** orientieren. Mit Blick auf gängige Standards für Evaluation muss also immer sichergestellt sein, dass durch einen KI-Einsatz die *Nützlichkeit*, *Fairness*, *Durchführbarkeit* und *Genauigkeit* von Evaluationen nicht negativ beeinflusst werden.

8 Letztlich maßgeblich sind immer die Standards guter Evaluation.

Mit Blick auf die Standards für Evaluation (DeGEval, 2017) ist beim KI-Einsatz besonders zu berücksichtigen:

- ***Genauigkeit***: Gute Evaluationen müssen verlässliche Ergebnisse bereitstellen. Von oder mit KI generierte Ergebnisse müssen also je nach Fall um angemessene Kontroll- und Absicherungsmaßnahmen ergänzt werden, um diese sicherzustellen. Die zuverlässige Replizierbarkeit von Ergebnissen ist dabei eine besondere Herausforderung.
- ***Fairness***: Gute Evaluationen sind fair und unparteiisch. Die oft dokumentierte Anfälligkeit von KI-Modellen gegenüber diskriminierenden *Biases*, also der einseitigen Interpretation und unausgewogenen Repräsentanz diverser Perspektiven, kann diese Fairness eklatant bedrohen.
- ***Durchführbarkeit***: Gute Evaluationen sind gegenstandsangemessen und praktikabel. Potenzielle Datenschutzverletzungen, zu denen es beim KI-Einsatz kommen kann, stellen dabei ein Hindernis dar. Weitere Hindernisse ergeben sich aus den für einen kompetenten KI-Einsatz erforderlichen Kompetenzen und Ressourcen.
- ***Nützlichkeit***: Das Gebot der Nützlichkeit von Evaluationen kann durch die starke Kontextabhängigkeit der Qualität von KI-Ergebnissen beeinträchtigt werden. Vor allem aber sind mögliche negative Auswirkungen auf das Beziehungsgefüge in Evaluationen zu berücksichtigen.

9 Die Entscheidung über den KI-Einsatz muss bei der Evaluation verbleiben.

Professionelle Evaluation zeichnet sich dadurch aus, dass sie unter Berücksichtigung einschlägiger professioneller Standards Evaluationskonzepte entwickelt und umsetzt, um damit vorhandene Informations- und Nutzungsbedarfe angemessen zu adressieren. Die Entscheidungen über Vorgehensweise und Methoden betreffen den ***Kern des professionellen Wissens*** von Evaluation. Nur wenn diese Entscheidungen bei den Evaluierenden verbleiben, kann die Neutralität der Evaluation gewährleistet und die Verantwortung für die Qualität der Ergebnisse eindeutig zugeschrieben werden.

Der KI-Einsatz ist hierbei keine Ausnahme. Da KI einen direkten Einfluss auf die Qualität von Evaluationsprozessen und –ergebnissen haben kann, muss die Entscheidung über ihre Nutzung in den Händen derer verbleiben, die diese Qualität in erster Linie zu verantworten haben und fachlich kompetent über den angemessenen Methodeneinsatz entscheiden können.