Bewertung der Datenqualität von Online-Umfragen: Eine explorative Untersuchung der Potenziale von Large-Language Models zur Detektion von unzureichendem Antwortaufwand.

Online-Umfragen sind in der empirischen Forschung als Instrument zur Datenerhebung weit verbreitet. Um Teilnehmer für Online-Umfragen zu erreichen, gewinnen Crowdsourcing-Plattformen wie Amazon Mechanical Turk oder Prolific zunehmend an Bedeutung. Solche Plattformen ermöglichen eine schnelle Datenerhebung und den Zugang zu oft diverseren Stichproben als herkömmliche Rekrutierungsmethoden. Es besteht jedoch die Gefahr, dass Teilnehmende hauptsächlich aus äusseren Anreizen wie beispielsweise der Vergütung teilnehmen und die Fragen daher unachtsam beantworten. Dieses als „Insufficient Effort Responding" (IER) bekannte Phänomen kann die Untersuchungsergebnisse erheblich verzerren.

Zur Erkennung und Kontrolle von IER wurden verschiedene Ansätze entwickelt. Beispielsweise testen instruierte Antworten, ob Teilnehmende Anweisungen befolgen, während statistische Verfahren Unstimmigkeiten in Antwortmustern aufdecken. Letztere erfordern jedoch häufig komplexe Analysen oder spezifische Datenvoraussetzungen. Besonders herausfordernd ist die Überprüfung offener Antworten, da diese meist manuell beurteilt werden müssen und stark vom jeweiligen Kontext abhängen, was besonders bei grösseren Stichproben erheblichen Aufwand bedeutet. Large Language Models (LLMs) können dabei unterstützen, doch ihre Zuverlässigkeit bei der Erkennung von IER ist noch unklar.

In diesem Workshop werden wir gemeinsam mit den Teilnehmenden explorieren, inwieweit sich LLMs zur Identifikation von IER eignen. Anhand öffentlich zugänglicher Datensätze sowie grösseren Datensätze aus unserer eigenen Forschung wollen wir gemeinsam erarbeiten, wie gut LLMs Vergleich zu menschlichen Einschätzungen und statistischen Kennwerten abschneiden.

Der Workshop beginnt mit einer kurzen Einführung in die Thematik der Datenqualität in Online-Untersuchungen und den Herausforderungen von IER. Wir stellen gängige Techniken vor, beleuchten deren Vor- und Nachteile und entwickeln gemeinsam erste Hypothesen zu den Potenzialen und Schwächen von LLMs in diesen Aufgaben. Anschliessend erhalten die Teilnehmenden einen Datensatz mit offenen Antworten aus einer früheren Studie, die bereits manuell bewertet wurden. In Gruppenarbeit erfolgt zunächst eine eigene Bewertung, die als zusätzliche Baseline dient. Im Anschluss vergleichen wir die aktuellen Einschätzungen mit früheren Ergebnissen, den Bewertungen eines LLMs und statistischen Kennzahlen. Ziel ist es, systematische Übereinstimmungen oder Abweichungen herauszuarbeiten.

Abschliessend folgt eine gemeinsame Reflexion darüber, unter welchen Bedingungen LLMs eine sinnvolle Ergänzung oder gar Alternative zu etablierten Verfahren darstellen könnten, wie sich mögliche Fehler systematisch erklären lassen und welche Prompt-Strategien zu besseren Ergebnissen führen könnten.

Ziel des Workshops ist es, praxisnah zu ermitteln, ob und unter welchen Bedingungen LLMs die Identifikation von IER sinnvoll unterstützen können. Dabei werden systematisch Stärken, Schwächen und Optimierungsmöglichkeiten herausgearbeitet.