Analyse qualitativer Daten mit KI-Unterstützung

Evolution statt Revolution qualitativer Evaluationsmethoden in der Qualitätssicherung der Philipps-Universität Marburg[1]

Florian Engel und Tristan Dohnt

Im Rahmen qualitativer Evaluationen hat sich das Einsatzspektrum von KI seit der Veröffentlichung von Large Language Models (LLMs) signifikant erweitert. Ein solches Einsatzszenario wird im Artikel vorgestellt: Die Einbindung der KI zur Unterstützung der inhaltsanalytischen Datenauswertung von Gruppendiskussionen in der Qualitätssicherungsarbeit von Studiengängen der Universität Marburg. Im Zentrum steht hierbei ein von uns entwickeltes, umfangreiches Codememo[2], das als Basis für die inhaltsanalytische Auswertung durch Künstliche Intelligenz dient. Eine knappe Reliabilitätsuntersuchung soll die Ergebnisgüte prüfen. Hierfür wurden zwei identische Datensätze sowohl von einem Menschen als auch von der KI inhaltsanalytisch kodiert/kategorisiert. Die Ergebnisse werden anschließend diskutiert. Es zeigt sich, dass es sich lohnt, bestehende Auswertungssysteme evolutionär weiterzuentwickeln, anstatt sie für den Einsatz durch KI von Grund auf neu zu konstruieren.

Analysis of qualitative data with AI support. Evolution instead of revolution of qualitative evaluation methods in quality assurance at Philipps University Marburg
The range of applications of AI in qualitative evaluations has expanded significantly since the publication of LargeLanguageModels. One such application scenario is presented in this article: the integration of AI to support the content-analytical data analysis of group discussions in the quality assurance work of degree programs at the University of Marburg. The focus here is a comprehensive code memo we developed, which serves as the basis for the content-

1 Basiert auf einem Impulsbeitrag im Rahmen der Tagung.

2 Codememos sind schriftlich festgehaltene analytische Notizen bzgl. eines Codes bzw. einer Kategorie (je nach Benennungskonvetion). Das Codememo dient dazu, die Bedeutung des/der übergeordneten Codes/Kategorie zwischen mehreren Analysierenden zu definieren und zu anderen Codes/Kategorien abzugrenzen. Sie dokumentieren damit die interpretativen Entscheidungen der Forschenden und unterstützen eine konsistente, transparente und nachvollziehbare Codierung über den gesamten Analyseverlauf hinweg – sind also eine gemeinsam entwickelte Konvention. (Saldana 2021 beschreibt coding memos, ohne sie selbst zu verwenden, Kukartz & Rädier 2024).

analytical analysis by artificial intelligence. A brief reliability study is intended to assess the quality of the results. For this purpose, two identical data sets were coded/categorized using content analysis by both a human and an AI. The results are then discussed. It shows that it is worthwhile to develop existing evaluation systems in an evolutionary manner rather than constructing them from scratch for use by AI.

Einleitung

Evaluationen gelten auch in Deutschland seit Jahrzehnten als zunehmend eigenes Forschungs- und Anwendungsfeld (Stockmann 2022). Gleichzeitig bezieht sich diese an sich bereits diverse Disziplin in Erkenntnistheorie und Methodologie weiterhin in großem Maße auf die ihr nahestehenden Geistes- und Sozialwissenschaften (Bryda et al. 2023). So gilt hier wie dort die Nutzung computergestützter qualitativer Datenanalyse (vermittels dezidierter QDA-Software) als Standard für Ausbildung und Alltagsgeschäft (Kuckartz/Rädiker 2024). Diese etablierten Vorgehensweisen sehen sich durch neue Entwicklungen im Bereich automatisierter Analyse ausgesetzt. An der Universität Marburg wollen wir diesen Wandel bewusst gestalten und unser methodisches Vorgehen sukzessive mit Hilfe von KI (weiter-)entwickeln. In diesem Artikel stellen wir einen Ansatz zur Integration von KI in die Analyse qualitativer Evaluationsdaten dar und beschreiben den damit einhergehenden Wandel als Evolution bestehender Methoden.

Eingangs erläutern wir unser generelles Vorgehen im Rahmen des Marburger Modells der Qualitätssicherung in Studium und Lehre. Anschließend gehen wir auf die qualitative Analyse von Evaluationsgesprächen ohne KI-Unterstützung ein und heben die Dokumentation in Form eines ›Kategorienschemas‹ hervor. Den Hauptteil beginnen wir mit einer knappen Erläuterung der beschafften technischen Lösung. Anschließend steht das sogenannte ›Codememo‹ als Herzstück unseres Ansatzes im Fokus. Hierbei handelt es sich um eine umfangreiche Sammlung an Kategoriendefinitionen und Beispielen, die unsere KI-unterstützte Analyse erst ermöglichen. Im abschließenden Abschnitt des Hauptteils weisen wir aus, wie wir die Ähnlichkeit der Kodierungen von Mensch und Maschine überprüft haben, was in das abschließende Fazit hinüberführt.

Qualitätssicherung und qualitative Inhaltsanalyse an der Universität Marburg

Als Teil der Zentralverwaltung ist es Aufgabe unseres Referats, Studiengänge über den gesamten student life cycle mit Evaluationsangeboten zu versorgen. Evaluation wird an der Universität Marburg in besonderem Maße fragestellungsbezogen be-

trieben (Stockmann 2002). Das heißt, es werden kaum regelmäßige universitätsweite Erhebungen durchgeführt. Stattdessen verfolgen wir ein Beratungsmodell, in dem jeder Qualitätssicherungskreislauf mit Sondierungsgesprächen beginnt (angelehnt an den PDCA-Zirkel (Deming 1986)). Hiervon ausgehend werden Daten in enger Abstimmung mit den relevanten Akteur*innen erhoben, ausgewertet und diskutiert, sowie darauf basierend Maßnahmen entwickelt, die in Zukunft überprüft werden sollen (Heinrich, 2020, Stockmann 2006).[3]

Nach der Klärung der interessierenden Themen und Fragen sowie der Entscheidung für einen qualitativen Workshop, setzen wir diesen in einer zwei- (QualiCheckFeedback) bis vierstündigen (QualiCheckStudientag) Session um, in der Audiodaten sowie Stichworte als Ergebnissicherung auf Postern als Dokumentation des Gesagten dienen. Diese werden im Sinne der qualitativen Inhaltsanalyse nach Kuckartz und Rädiker (2024) analysiert.

Qualitative Inhaltsanalyse bedeutet, Daten anhand deduktiv und/oder induktiv gebildeter Hauptkategorien sukzessive und immer feiner zu analysieren. Zentral für das Vorgehen ist, dass bereits absolvierte Analyseschritte erneut durchgeführt werden, wenn neue Daten vorliegen oder sich das Kategoriensystem derart weiterentwickelt hat, dass eine neuerliche Überprüfung von Kategorien durch erneutes Codieren sinnvoll erscheint.

Im Gegensatz zur Methodenumsetzung mit wissenschaftlichem Anspruch, vor allem eines iterativen Auswertungsprozesses, besteht die Datenerhebung und -auswertung an der Universität Marburg jeweils aus einem zusammenhängenden Schritt und dauert in aller Regel höchstens sechs Monate von der Instrumentenabsprache (über die Bewerbung und Durchführung des Workshops) bis zum fertigen Bericht. Gerade der Aspekt der Überprüfung von Daten und des Zurücktretens im Prozess – umhin also die langfristige Weiterentwicklung und Verbesserung der Kategorien, ist in unserer alltäglichen Arbeit nur eingeschränkt möglich.

Als Antwort darauf haben wir vor einer Weile begonnen, die reichlich vorhandenen Berichte vergangener qualitativer Evaluationsmaßnahmen zu sammeln, zu sortieren und zu kategorisieren. Das Ergebnis ist eine sehr umfangreiche Excel-Tabelle, die uns als Dokumentation und Erinnerungsanker dafür dient, welche Haupt- und Subkategorien aus welchen Daten über die Jahre entwickelt wurden. Mit dieser als ›Kategorienschema‹ benannten Datei lassen sich neu erhobene Datensätze in Bezug zu vergangenen Evaluationsdurchläufen setzen und auswerten. Durch die fortlaufende Implementation neuer Auswertungsberichte aktualisieren wir zugleich diese Datenbasis des ›Kategorienschemas‹ und erfüllen damit ›in the long run‹ den Anspruch des kuckartz'schen Vorgehens, in der Analyse an bereits erfolgte

3 Für eine ausführliche, aktuelle Beschreibung des Marburger Ansatzes siehe Dohnt/Engel 2025.

Arbeitsschritte zurückzukehren (Kuckartz/Rädiker 2024: 142). Tatsächlich überprüfen wir damit nicht nur neue Daten, sondern modifizieren und verbessern unser Kategoriensystem fortwährend.

Qualitative Analyse von Evaluationsdaten mit generativer KI

Technische Basis

Für die Umsetzung unserer KI-Analysen haben wir das Modul AI Assist für das Programm MaxQDA der Firma Verbi beschafft. Dieses nutzt zum Zeitpunkt der Studie im Juli 2025 das LLM ChatGPT von OpenAI in der Version 3.5. AI Assist bietet verschiedene Funktionen zur Nutzung generativer KI im Arbeitsprozess. Im Rahmen erster Tests erwies sich die Funktion mit multiplen Dokumenten zu chatten als sinnvollste Variante. Während andere Funktionen einzelne Kategorien automatisch zusammenfassen oder analysieren, wird im Dokumenten-Chat ein separates Chatfenster im User*inneninterface geöffnet, über das mit allen ausgewählten Dateien (auch dem Codememo) zugleich gearbeitet werden kann, was sehr ähnlich zur Bedienung in Schnittstellen via Browser funktioniert. Eine hier angewiesene, automatische Analyse wird so verarbeitet, als würde es sich um ein einzelnes Dokument handeln. Zugleich besteht keine Möglichkeit, weitere Einstellungen bzgl. Kreativität, Logik, Fokus oder Diversität der Antworten der eingesetzten KI vorzunehmen, wie in gängigen Browserschnittstellen – was angesichts der Nachvollzieh- und Reproduzierbarkeit durchaus sinnvoll erscheint.

Codememo

Von zentraler Bedeutung unseres Ansatzes ist – wie bereits angedeutet – das sogenannte ›Codememo‹. Ein Codememo stellt in der Anwendung von QDA-Software durch einen Menschen eine kurze Zusammenfassung dessen dar, was eine Kategorie bedeuten und umfassen soll und ist im User*inneninterface auch mit diesem verbunden (Kuckartz/Rädiker 2024: 122–123, 208–214). Dies ermöglicht es menschlichen Anwender*innen, ein gemeinsames Verständnis einer (Sub-)Kategorie zu entwickeln, was die Analyse und ihre Ergebnisse besser und nachvollziehbarer gestaltet. Auch ›unser‹ Codememo versammelt Definitionen und Beispiele für existierende Haupt- und Subkategorien, die wir aus der deutlich umfangreicheren Kategorienschema-Datei entwickelt haben. Diese Excel-Datei wurde hierzu sowohl händisch wie auch durch KI zusammengefasst, strukturiert und neu formatiert. Es umfasst in seiner aktuellen Version 12 Oberkategorien, 92 Unterkategorien sowie knappe Analysehinweise genereller Art mit insgesamt knapp 65.000 Zeichen in ei-

ner einzelnen Word-Datei. Diese Datei durfte aufgrund der Zeichenbegrenzungen von AI Assist im Juli 2025 100.000 Zeichen nicht überschreiten.

Die Definitionen jeder Hauptkategorie werden im Codememo nach demselben Muster festgehalten:

1. Definitionen Oberkategorie (erläuternder Fließtext plus Beispiele)
2. Definitionen Subkategorien (erläuternder Text plus Beispiele in bullet point-Struktur)
3. Analysehinweise genereller Art (erläuternder Text in bullet point-Struktur)

Dieses Codememo erfüllt also einen vergleichbaren Zweck wie das gleichnamige Element im gängigen analytischen Vorgehen menschlicher Evaluator*innen und Sozialforscher*innen (Kuckartz/Rädiker 2024: 208–214). Statt eine Konvention unter menschlichen Analyst*innen zu sein, liefert es einer generativen KI die grundlegenden Informationen dazu, anhand welcher Kategorien und Definitionen »sie« vorgegebene Daten zu analysieren hat. Der genutzte Prompt enthält neben basalen Verarbeitungsanweisungen auch Kontextinformationen zu den Daten, den zentralen Verarbeitungsauftrag sowie Ausgabeanweisungen. Letztere dienen dazu, eine stabile, in unsere Berichtsform übertragbare Ergebnisstruktur zu erhalten. In Zukunft sollen auch diese Ergebnisse zur Überarbeitung und Weiterentwicklung des Kategorienschemas und damit des Codememos herangezogen werden.

Vergleich der Kodierungen von Mensch und Maschine

Im Rahmen des Entwicklungsprozesses haben wir Daten aus zwei verschiedenen Evaluationsmaßnahmen (einer Modul- und einer Studiengangsevaluation) sowohl von einem Menschen als auch mit unserem KI-Verfahren inklusive Codememo auswerten lassen. Die Analyse umfasste insgesamt 75 Zitatstellen. Mensch und KI ordneten in 74,14 % der Fälle die Zitate derselben Oberkategorie und in 87,88 % der Fälle derselben Unterkategorie zu. Somit war die Übereinstimmung zwischen Mensch und KI signifikant höher als der Zufall (Oberkategorie: $t(57)=4{,}16$, $p<.001$; Unterkategorie: $t(65)=9{,}36$, $p<.001$)[4]. Die Kodierungen wurden von zwei unabhängigen Raterinnen durchgeführt. Die Interrater-Reliabilität lag bei der Oberkategorie bei Cohen's Kappa $\kappa = 0{,}486$ und bei der Unterkategorie bei Cohen's Kappa $\kappa = 0{,}575$, was eine mäßige, aber signifikante Übereinstimmung zwischen den Raterinnen darstellt und somit als verlässlich angesehen werden kann.

Für den ersten Test sind wir mit diesen Ergebnissen zufrieden. Sie legen nahe, dass die entwickelte automatische Auswertungsmethode Ergebnisse liefert, die mit

4 Für die Vergleiche und Signifikanztests wurden nur die Fälle berücksichtigt, in denen die beiden Raterinnen in ihren Urteilen übereingestimmt haben.

Ergebnissen unserer eigenen Analysen vergleichbar genug sind, um als Basis für die künftige Weiterentwicklung zu dienen, ähnlich den Erkenntnissen von Gao et al. (2023).

Fazit

Wie im letzten Abschnitt zum Reliabilitätsmaß bereits angedeutet wurde, haben sich das Codememo und die Vorgehensweise der Auswertung bis zu diesem Zeitpunkt bereits bewährt. Der unausweichlichen Voreingenommenheit menschlicher Rater*innen wurde zwar durch Beispiele für das Rating seitens der Autor*innen begegnet – menschliche Bewertung bleibt letzten Endes aber ein subjektiver Prozess. »Opak« (Schmitt/Heckwolf 2024: 53) bleibt aber auch, was im Prozess der »tatsächliche[n] Verarbeitung der Daten« einer mit Deep Learning trainierten KI geschieht. Was jedoch bereits an der Oberfläche von KI produzierten Textes auffällt und sich auch in der Literatur findet, ist, dass KI-Analyse (weiterhin) Kontextinformationen fehlen bzw. sie diese nicht (sicher) erfasst, was Ergebnisse oberflächlicher Zuordnungen weniger nachvollziehbar und Schlussfolgerungen teils widersprüchlich wirken lässt (Roberts et al. 2023; Mostafapour et al. 2024; Thelwall 2025).

Verbesserungen erscheinen dennoch möglich. Diese konzentrieren sich derzeit vornehmlich auf das Codememo als zentrales analytisches Raster. Dieses sollte vor der nächsten Anwendung um die aktuellen Ergebnisse ergänzt werden. Gerade hier fehlt uns aber noch ein erprobtes Vorgehen, das nicht den gesamten Entwicklungsprozess neu aufrollt. Schon aus dieser Sicht handelt es sich bei unserem Ansatz um einen evolutionären, insofern wir neue Prozesse entlang des Weges entwickeln und bei Erfolg implementieren. Verbesserung ist aber auch im Prozessablauf nötig und möglich. Andere Untersuchungen verweisen zusätzlich darauf, dass Prompts ein beständiges Feld der Unsicherheit und Verbesserung darstellen (Fuller et al. 2024; Mostafapour et al. 2024) – was wir einerseits bestätigen können, andererseits nicht die zentrale Herausforderung darstellt.

Das analytische Vorgehen – mit dem existierenden Codeschema – innerhalb des QDA-Programms ist dagegen derart routinisiert, dass es sich bei kommenden Evaluationsmaßnahmen bereits anbietet, menschliche und automatisch erstellte Auswertung quasi parallel durchzuführen, inhaltlich zu vergleichen und die menschliche Version durch die KI-erstellte zu ergänzen. Ideal wäre es allerdings zugleich, erneute Reliabilitätsüberprüfungen in den Ablauf zu integrieren. Hieran erkennt man wiederum die Limitationen einer Entwicklung entlang des Weges, da sich Überprüfungen und Tests nicht immer in das Alltagsgeschäft einbinden lassen.

Literatur

Bryda, G./Costa, A. P. (2023): Qualitative Research in Digital Era: Innovations, Methodologies and Collaborations, in: Social Sciences, 12 (10), 570.

Deming, W. E. (1986): Out of the Crisis, Cambridge: MIT Press.

Dohnt, Tristan/Engel, Florian (2025): »Qualitätssicherung als per se ko-kreativer Prozess. Die Marburger Qualitätssicherung für Studium und Lehre«. in: Jörg Jörissen/Daniela Möller/Martina Grein/Sacha Kopczynski/David Peters (Hg.), Evaluation von Studium und Lehre ko-kreativ gestalten, Münster/New York: Waxmann, S. 101–108.

Heinrich, Mathis (2020): »Nachhaltigkeit durch Flexibilität, Reichweite und Diffusion: Das Marburger Modell einer Qualitätssicherung »on demand«. in: QiW – Qualität in der Wissenschaft, 14(1), 17–26.

Kuckartz, Udo/Rädiker, Stefan (2024): Qualitative Inhaltsanalyse. Methoden, Praxis, Umsetzung mit Software und künstlicher Intelligenz, Weinheim: Beltz Juventa.

Roberts, Richard Hr./Ali, Stephen R./Hutchings, Hayley A./Dobbs, Thomas D./Whitaker, Ian S. (2023): »Comparative study of ChatGPT and human evaluators on the assessment of medical literature according to recognised reporting standards«, in: BMJ Health and Care Informatics, 30 (1).

Saldana, Johnny (2021): The Coding Manual for Qualitative Researchers, Los Angeles/London/New Dehli/Singapore/Washington DC: SAGE

Schmitt, Marco/Heckwolf, Christoph (2024): »KI zwischen Blackbox und Transparenz:«, in: Roger Häußling/Claudius Härpfer/Marco Schmitt (Hg), Soziologie der Künstlichen Intelligenz: Perspektiven der Relationalen Soziologie und Netzwerkforschung, Bielefeld: transcript Verlag, S. 51–84.

Stockmann, Reinhard (2002): Was ist eine gute Evaluation. CEval-Arbeitspapiere, Nr. 9, Centrum für Evaluation. PID: https://nbn-resolving.org/urn:nbn:de:0168-ssoar-118018

Stockmann, Reinhard (2006): Evaluation und Qualitätsentwicklung. Eine Grundlage für wirkungsorientiertes Qualitätsmanagement. (Bd. 5), Münster/New York: Waxmann.

Stockmann, Reinhard (Hg.) (2022): Handbuch zur Evaluation. Eine praktische Handlungsanleitung. (Band 16), Münster/New York: Waxmann.

Thelwall, Mike (2025): »ChatGPT for complex text evaluation tasks«, in: Journal of the Association for Information Science and Technology, 76 (4), S. 645-648.