

Qualitative Inhaltsanalyse mit ChatGPT: Fallstricke, grobe Annäherungen und grobe Fehler. Ein Erfahrungsbericht

Philipp Mayring

Keywords:

ChatGPT;
künstliche
Intelligenz;
qualitative
Inhaltsanalyse;
Themenanalyse;
qualitative
Textanalyse;
Grounded-Theory-
Methodologie;
psychoanalytische
Textinterpretation

Zusammenfassung: In diesem Beitrag wird durch eine Reihe von Testläufen untersucht, welchen Beitrag das KI-basierte Programm ChatGPT in den beiden Versionen 3.5. und 4 zu einer qualitativen Inhaltsanalyse von Interviewtexten leisten kann. Dazu wird ein kurzer Mustertext mit meiner Musterlösung präsentiert. Grobe Eingaben bei einer eher naiven Nutzung ("Führe eine qualitative Inhaltsanalyse durch!") als auch differenzierte Vorgaben mit Fragestellung und genaueren Kodieranweisungen führten in beiden Versionen höchstens zu groben Annäherungen an die Musterlösung bei einer Vielzahl von groben Fehlern. Das Programm hatte auf unterschiedliche inhaltsanalytische Konzepte (BRAUN & CLARKE 2006; KUCKARTZ 2014; MAYRING 2022a; SCHREIER 2012) nicht bzw. falsch reagiert, erkannte verborgene Textinhalte nicht und versagte bei Überprüfungen der Kodierübereinstimmung. Die Ergebnisse der Software wiesen, egal welche Vorgaben gemacht wurden, meistens in Richtung einer groben, oberflächlichen Zusammenfassung im Sinne einer Themenaufzählung. Für die von mir entwickelten qualitativ-inhaltsanalytischen Verfahren (MAYRING 2022a, 2022b) erscheinen sie wenig geeignet.

Inhaltsverzeichnis

- [1. Einleitung](#)
 - [2. Das Beispielmateriale](#)
 - [3. Qualitative Inhaltsanalyse mit ChatGPT 3.5](#)
 - [4. Präzisierung der Eingabe](#)
 - [5. Kodierübereinstimmung](#)
 - [6. Größere Materialmengen mit der kostenpflichtigen Version ChatGPT 4](#)
 - [7. Fazit](#)
- [Literatur](#)
- [Zum Autor](#)
- [Zitation](#)

1. Einleitung

Kürzlich erlebte ich während einer universitären Veranstaltung über künstliche Intelligenz (KI) in Forschung und Lehre Folgendes: Ein Dozent berichtete, dass er seinen Studierenden für Bachelor- und Masterarbeiten gerne rate, wenn sie Interviews für die Datenerhebung durchgeführt hatten, das Material erst einmal in ChatGPT einzugeben und die KI zu bitten, eine qualitative Inhaltsanalyse nach MAYRING (2022b) durchzuführen. Ich war perplex. Über 40 Jahre haben mein Team und ich Techniken qualitativer Inhaltsanalyse mit Ablaufmodellen und prozeduralen Regeln immer weiter verfeinert, eine interaktive Software ([QCAmap](#)) wurde entwickelt, die dies unterstützt, und nun soll ChatGPT das Ganze automatisch in Sekundenschnelle genauso ausführen ohne Lehrbuchstudium, ohne Beratung, ohne Einführungsworkshop? [1]

Um die Leistungen von ChatGPT zu überprüfen, stelle ich einen Beispieltext vor sowie eine in unserer Arbeitsgruppe entwickelte Musterlösung einer qualitativen Inhaltsanalyse (Abschnitt 2). In einer ganzen Reihe von Versuchen forderte ich eine Inhaltsanalyse von ChatGPT an (Abschnitt 3), auch mit präzisen *Prompt*-Eingaben (Abschnitt 4) sowie größeren Materialmengen (Abschnitt 6) und vergleiche die Ergebnisse mit unserer Musterlösung. Das Problem mangelnder Kodierübereinstimmung bei erneuter Eingabe wird in Abschnitt 5 diskutiert und am Ende ein eher ernüchterndes Fazit gezogen (Abschnitt 7). [2]

2. Das Beispielmateriale

Zuhause am Schreibtisch setzte ich mich sofort hin und probierte es aus. Ich verwendete einen Text, den ich gerne in Workshops als Anschauungsmaterial einsetze. Der Ausschnitt stammt aus einem Interview mit einem ehemaligen Lehrer in Ostdeutschland, der nach der Wende arbeitslos geworden ist (MAYRING, KÖNIG, BIRK & HURST 2000).

"P¹: Das erste Mal wurde mir als Berufsschullehrer gekündigt, weil ich neben vielen anderen Fächern ein Fach unterrichtete, Staatsbürgerkunde, was heute in der Stundentafel des Landes Sachsen nicht mehr enthalten ist. Ich habe dann nach fünfmonatiger Arbeitslosigkeit 10 Monate lang in einer privaten Bildungsanstalt Umschulung arbeitsloser Werktätiger gearbeitet und habe dort Fächer wie allgemeine Wirtschaftslehre, Betriebswirtschaftslehre, kaufmännisches Rechnen, Buchführung und alles, was anfiel, unterrichtet. Aber diese Firmen kriegen im Moment vom Arbeitsamt keine Zuweisung mehr, sodass ich seit April zum zweiten Mal arbeitslos bin.

I: Ist diese Arbeitslosensituation im Augenblick für Sie belastend?

P: Ja das ist schon belastend, belastend in der Gestalt, dass man sich mit anderen Kollegen vergleicht, die noch im Dienst sind und man selber weiß, was man wert ist und was man kann und was man nicht kann. Und ich glaube, dass hier viele Entlassungen in die Arbeitslosigkeit ganz einfach willkürlich vorgenommen worden sind und dort jegliche Gerechtigkeit am Boden geblieben ist. Das hab ich auch schon dreimal vor dem Arbeitsgericht versucht, mein Recht zu bekommen und ich hoffe, dass mir das im vierten Anlauf gelingt.

I: Hat die Arbeitslosigkeit auch positive Seite für Sie?

P: Eigentlich würde ich das verneinen, weil ich sehr, sehr gerne Lehrer war, sehr, sehr gerne unterrichtet habe und alleine nur die Wohnung in Ordnung bringen, den Garten in Ordnung bringen oder einzukaufen, das füllt mich insgesamt gesehen nicht aus.

I: Glauben Sie, dass Sie mit den Belastungen fertig werden?

P: Ja.

I: Was haben Sie, Sie haben's schon angesprochen, mit dem Einklagen beim Arbeitsgericht, haben Sie ansonsten was unternommen gegen die Situation?

¹ P steht für Person, I steht für Interviewer. Das Interview wurde nicht wörtlich im Buch abgedruckt, stammt aber aus dem erwähnten Projekt.

P: Na ich hab ja alle Möglichkeiten über die Gewerkschaft Erziehung und Wissenschaft genutzt, die mir da geboten werden und das Komplizierte dabei ist, dass in dem Prozess nach der Wende viele Dinge, was Rechtsbewusstsein ist, was Arbeit mit dem Betriebsrat ist und viele andere Dinge mehr, womit der Bundesbürger in 40 Jahren Bundesrepublik groß geworden ist, durch Elternhaus, durch Schule, durch Erziehung, durch Umfeld, alles das musste der DDR-Bürger, Ex-DDR-Bürger, in den letzten drei oder vier Jahren zusätzlich zu seiner täglichen Arbeit, zu seinem Lebensablauf lernen. Damit glaube ich, dass ich die Möglichkeiten, die mir bekannt waren, genutzt habe, sowohl mich weiterzubilden, u.a. bei der Bundeszentrale für politische Bildung in M, zum Weiterbildungslehrgang von einer Woche, habe auch ansonsten alle Möglichkeiten der Literatur usw. usw. genutzt.

I: Würden Sie denn für Ihre gesamte psychische Situation eine positive Bilanz ziehen?

P: Ja, ich würde sagen ja.

I: Und die einzigen Belastungen, die da sind, sind dann bedingt durch die Gänge zum Arbeitsamt?

P: Durch die Gänge zum Arbeitsamt und ganz einfach deshalb, weil ich noch nicht weiß, wie das Arbeitsgerichtsverfahren entschieden wird. Weil ich auch nicht weiß, wie lange meine Frau noch Arbeit haben wird und wenn ich natürlich kein Arbeitslosengeld mehr kriegen würde und meine Frau wäre arbeitslos, dann wär's schon echt belastender.

I: Also in erster Linie ist im Augenblick die Unsicherheit der Situation.

P: Ja." [3]

Ich möchte zunächst meine Auswertung vorstellen. Ich beziehe mich dabei auf die Konzeption qualitativer Inhaltsanalyse nach MAYRING (2022a,2022b) sowie MAYRING und FENZL (2022). Ziel ist dabei, im Gegensatz zu anderen Konzeptionen (z.B. KUCKARTZ 2014, SCHREIER 2012, s. Abschnitt 3) eine möglichst exakte, per Kodierübereinstimmung überprüfbare Kategorienzuordnung vorzunehmen. Dazu sind für jede der einzelnen vorgeschlagenen Techniken (Zusammenfassung, induktive Kategorienbildung, Explikation, deduktive Kategorienanwendung, Mischtechniken) klare Verfahrensschritte und Ablaufmodelle entwickelt worden, sodass es Sinn macht, von einer Musterlösung zu sprechen. [4]

Die Fragestellung der Analyse war: Welche Belastungsfaktoren durch die Arbeitslosigkeit wurden berichtet? Ich bestimmte danach die anzuwendende Technik als induktive Kategorienbildung. Dafür sind als inhaltsanalytische Regeln die Kategoriendefinition sowie das Abstraktionsniveau festzulegen. Die Kategoriendefinition lautete "Subjektiv negativ erlebte Aspekte der Veränderungen der Lebenssituation durch die Arbeitslosigkeit". Das erwünschte Abstraktionsniveau war eine möglichst konkrete Kategorienformulierung am Text, jedoch nicht so spezifisch, dass die Kategorie nur bei der betreffenden Person kodiert werden kann. [5]

Nach diesen Regeln ergeben sich sechs induktive Kategorien:

- K1: Willkürliche, ungerechte Kündigung
- K2: Verlust des geliebten Lehrerberufs
- K3: Haushalt und Garten füllen nicht aus
- K4: Unvorbereitet auf neue Rahmenbedingungen nach der Wende
- K5: Unsicherheit über Ausgang der Klage gegen die Entlassung
- K6: Finanzielle Unsicherheit in der Zukunft [6]

"Behördengänge" ist nicht kategorisiert worden, da es nach übereinstimmendem Urteil der AuswerterInnen nur auf eine Suggestivfrage genannt wurde und gleich darauf relativiert wurde. Man "weiß, was man wert ist und was man kann" wurde nicht als Verletzung des Selbstwertgefühls kategorisiert, da dies als im Vergleich zum formulierten Abstraktionsniveau und der Kategoriendefinition zu interpretativ gewesen wäre. [7]

3. Qualitative Inhaltsanalyse mit ChatGPT 3.5

Um [ChatGPT](#) zu testen, verwendete ich zunächst die kostenlos zugängliche Version 3.5 und begann, dem "Rat" des in der Einleitung zitierten Kollegen folgend, ganz offen: "Mach mir bitte eine Qualitative Inhaltsanalyse nach Mayring mit folgendem Interviewausschnitt." Dies ist natürlich eine völlig unzureichende, naive Anfrage, die aber oft vorkommen dürfte. Insofern stellt dies einen Fallstrick des Arbeitens mit ChatGPT dar. Auf die obige Frage kam sehr rasch die Antwort (ChatGPT 3.5 vom 24. März 2024):

"Um eine qualitative Inhaltsanalyse nach Mayring durchzuführen, werden wir den Interviewausschnitt inhaltlich analysieren und in Kategorien zusammenfassen. Hier sind die Schritte für die Analyse:

Schritt 1: Transkription des Interviews ...

Schritt 2: Kategorienbildung

Gründe für Arbeitslosigkeit:

Staatsbürgerkunde nicht mehr im Lehrplan

Mangel an Zuweisungen seitens des Arbeitsamtes

Belastung durch Arbeitslosigkeit:

Vergleich mit anderen Kollegen

Empfindung von Ungerechtigkeit

Rechtliche Auseinandersetzung vor dem Arbeitsgericht

Negative Auswirkungen der Arbeitslosigkeit:

Verlust des Lehrerberufs

Mangel an Erfüllung durch Hausarbeit

Bewältigung der Situation:

Glaube an die eigene Fähigkeit, mit der Belastung fertig zu werden

Nutzung von Gewerkschaftsdiensten und Weiterbildungsmöglichkeiten

Positive Aspekte der Situation:

Psychische Bilanz ist insgesamt positiv

Unsicherheit:

Unsicherheit bezüglich des Arbeitsgerichtsverfahrens

Unsicherheit über die Zukunft der Ehefrau

Schritt 3: Codieren des Materials: Das Interview wird entsprechend den identifizierten Kategorien codiert.

Schritt 4: Interpretation der Ergebnisse: Die Häufigkeit der Codes wird analysiert, um Muster und Trends zu identifizieren. Die Interpretation erfolgt im Kontext der Forschungsfrage oder des Ziels der Studie." [8]

Die Darstellung der Methode mit den vier angeführten Schritten ist fehlerhaft, in meinen Arbeiten so nirgends publiziert und ähnelt eher dem von SCHREIER (2012) vorgeschlagenen und von mir (MAYRING 2019a) als sehr aufwändig kritisierten Verfahren: erst die Kategorienentwicklung, dann in einem zweiten Materialdurchgang die Codierung. Das von mir und meinem Team entwickelte Vorgehen besteht hingegen aus acht verschiedenen Techniken, aus denen gemäß der Fragestellung ausgewählt werden muss. Entscheidet man sich für die induktive Kategorienbildung, muss nach der Festlegung der inhaltsanalytischen Einheiten (Kodiereinheit, Kontexteinheit, Auswertungseinheit) fragestellungsbezogen und theoriegeleitet eine Kategoriendefinition sowie ein Abstraktionsniveau festgelegt werden (siehe Abschnitt 2). Wenn dann Kategorien entwickelt wurden, wird die Möglichkeit angeboten, wiederum theoriegeleitet Überkategorien zu entwickeln, und nicht umgekehrt. Es wird vom Material ausgegangen und in der Folge abstrahiert. [9]

ChatGPT geht anscheinend eher von einer zusammenfassenden qualitativen Inhaltsanalyse aus, hält sich aber nicht an hierfür von mir (MAYRING 2022a, 2022b) vorgeschlagene Verfahrensschritte für Zusammenfassungen (Paraphrasierung, Reduktion, Generalisierung, weitere Reduktion[en] und Generalisierung[en]). Sieht man sich die beiden negativen Kategorien an, so werden in den Nennungen vier der sechs Kategorien aus der Musterlösung in etwa gefunden, bei drei Fehlkategorisierungen ein eher schlechtes Ergebnis. Eine Schlussfolgerung meiner Analyse (siehe Abschnitt 7) deutet sich bereits an: grobe Annäherungen und grobe Fehler. [10]

In einem weiteren Test habe ich die gleiche Frage mit der Bitte um Auswertung nach der qualitativen Inhaltsanalyse nach SCHREIER (2012) gestellt. In der Antwort (ChatGPT 3.5 vom 10. März 2024) wurden drei Schritte präsentiert, nämlich offenes Kodieren, axiales Kodieren und selektives Kodieren, die eigentlich aus der Grounded-Theory-Methodologie (GTM, GLASER & STRAUSS 1967; STRAUSS 1991 [1987]) stammen, von ChatGPT 3.5 aber falsch beschrieben und falsch durchgeführt wurden. Nach der GTM sollten aus dem offenen Codieren solche Codes für das axiale Codieren ausgewählt werden, die für die Theoriebildung ertragreich erscheinen. ChatGPT konnte das nicht

umsetzen. Die GTM wurde von SCHREIER (2012) wohl angeführt, aber nur als ein möglicher Schritt (neben Zusammenfassung, Subsumption und Kontrastierung) in der Kategorisierung des Materials. Auch hier könnte man einwenden, der Sinn des Arbeitens mit ChatGPT sei nicht das Stellen unspezifischer globaler Fragen, sondern das iterative Annähern durch Präzisierungen in der Fragengestaltung (*Prompts*) (LIEDER & SCHÄFFER 2024). Dazu müsste man aber einen klaren Begriff von der Methode und der spezifischen Fragestellung haben, was von vielen Nutzerinnen und Nutzern nicht zu erwarten sein dürfte. Hätte man die präzisen Vorstellungen schon zu Beginn, um daraus sinnvolle *Prompts* nachschicken zu können, könnte man die Analyse ja gleich direkt am Material ohne KI vornehmen, also wieder ein Fallstrick der ChatGPT-Nutzung. Ein iterativer Prozess der Annäherung an eine offene Fragestellung durch Verfeinerung der *Prompts* ist beim Arbeiten mit der qualitativen Inhaltsanalyse nach MAYRING (2022a) eigentlich nicht angesagt, da klare Fragestellungen und Kodierregeln vorab vorliegen. Bei offeneren, explorativen Ansätzen qualitativer Forschung mag dies anders sein. [11]

Die Antwort auf die gleiche Frage mit der Bitte um qualitativ-inhaltsanalytische Auswertung nach KUCKARTZ (2012, 2014) war ebenso unbefriedigend. Dieser hatte drei Basismethoden vorgeschlagen, die inhaltlich strukturierende, die evaluative (deduktive!) und die typenbildende qualitative Inhaltsanalyse. Das von ChatGPT 3.5 vom 10. März 2024 präsentierte Ergebnis ähnelte am ehesten der ersten induktiven Technik, was doch erklärt werden sollte. Die einzelnen Kategorien und untergeordnete Nennungen waren ähnlich der Lösung nach SCHREIER (2012), ebenso der Ablauf der Kategorienbildung und die nachfolgende Kodierung. [12]

Der gleiche Versuch mit der Bitte um eine thematische Analyse nach BRAUN und CLARKE (2006, 2024) ergab große Ähnlichkeiten zu den für KUCKARTZ (2012, 2014) ermittelten Ergebnissen, obwohl das Vorgehen von BRAUN und CLARKE zuvor in der ChatGPT-Antwort als auf verborgene Muster und Themen im Text abzielend charakterisiert worden war. Demgegenüber unterschieden BRAUN und CLARKE (2006, 2024) verschieden induktive und deduktive Formen der thematischen Analyse. Das Ziel, verborgene Muster zu identifizieren entspricht am ehesten der von ihnen beschriebene *Reflexive Theme Analysis*. Die von ChatGPT 3.5 präsentierten Ergebnisse enthielten jedoch keine verborgenen Muster, sondern blieben im inhaltlich Beschreibenden. [13]

Bei weiteren Tests war mir ein Fehler unterlaufen und ich hatte nach einer thematischen Analyse nach BROWN und HARRIS statt nach BRAUN und CLARKE gefragt. BROWN und HARRIS haben sich nie mit Textanalyse befasst, sind dagegen bekannt für psychologische Lebensereignis- und Depressionsforschung (BROWN & HARRIS 1978). Chat GPT führte trotzdem in beiden Versionen 3.5 und 4.0 eine Analyse durch und präsentierte mehr oder weniger sinnlose Ergebnisse, listete zum Beispiel Lebensereignisse auf, die im Text gar nicht genannt worden waren. [14]

Weitere Versuche mit ChatGPT ergaben, dass ironische Bemerkungen und indirekte Andeutungen nicht erkannt wurden. Auch der Versuch mit einer Textstelle, die einen Versprecher beinhaltet und aus psychoanalytischer Sicht als Indikator für Verdrängungsprozesse gedeutet werden kann (vgl. zu psychoanalytischen Textinterpretation HAUBL und LOHL 2020), wurde von ChatGPT nicht erkannt. Komplexere Auswertungsbefehle (z.B. "Mache mir bitte eine Liste der Kategorien mit den Häufigkeiten der Textstellen, die jeweils dazu gefunden wurden") wurden nicht berücksichtigt. Bei der Eingabe mehrerer Interviews stieß ich schnell an eine Grenze, die in etwa bei drei Transkriptseiten liegt, was für inhaltsanalytische Studien, die oft mehrere Hundert Seiten umfassen, ein Ausschlusskriterium darstellt. [15]

4. Präzisierung der Eingabe

Ein Grund für die unbefriedigenden Ergebnisse liegt bestimmt darin, dass die Analyseanfrage zu global war. Deshalb startete ich, dem Musterbeispiel folgend, den nächsten Versuch mit der präziseren Eingabe:

"Welche Belastungsfaktoren durch die Arbeitslosigkeit nennt die Person? Verwende bitte die induktive Kategorienbildung nach Mayring. Formuliere die Kategorien möglichst konkret am Text, aber nicht so spezifisch, dass sie nur bei dieser einen Person auftauchen würden. Kategorisiere nur, wenn der negative Faktor auch subjektiv bedeutsam ist". [16]

Dies entspricht den im Musterbeispiel formulierten Regeln. Die Antwort von ChatGPT 3.5 (vom 24. März 2024) war erheblich schlechter im Vergleich zur Musterlösung:

1. Vergleich mit anderen Kollegen
2. Empfindung von Ungerechtigkeit bei Entlassungen
3. Rechtliche Auseinandersetzung vor dem Arbeitsgericht
4. Unsicherheit bezüglich des Ausgangs des Arbeitsgerichtsverfahrens
5. Unsicherheit über die Zukunft der Ehefrau [17]

Die erste Nennung ist kein eigenständiger Belastungsfaktor, sondern die Ungerechtigkeit aufgrund des Vergleiches ist belastend. Die dritte Nennung ist ebenso keine eigenständige Belastung, sondern Ausdruck der Unsicherheit über den Ausgang des Verfahrens. Drei der Kategorien aus der Musterlösung ("Verlust des geliebten Lehrerberufs", "Haushalt und Garten füllen nicht aus", "Unvorbereitet auf neue Rahmenbedingungen nach der Wende"; siehe Abschnitt 2) wurden nicht erkannt. [18]

Nun könnte man die Eingaben weiter präzisieren, genaue Vorgaben und Erklärungen zu einzelnen Ablaufschritten, sogar Beispielanalysen in *Prompts* einfügen. Hier existieren auch bereits frei verfügbare Vorlagen, sogenannte *Custom GPTs*, und auch ein [Mayring Mentor](#) mit über Tausend Nutzerinnen und

Nutzern.² Ich rief die Vorlage auf und sah immerhin die Angabe eines Autors, der sich als Experte bezeichnete, aber laut Internetbeiträgen in der Bau- und Immobilienbranche tätig ist. Die Vorlage wies gravierende Fehler auf: Es wurde nur eine der von uns entwickelten Techniken beschrieben, nämlich die induktive Kategorienbildung. Diese wurde falsch charakterisiert, eine Paraphrasierung gefordert, die wir nur bei Zusammenfassungen vornehmen, nur eine Analyseeinheit angegeben (nur die Kodiereinheit, nicht die ebenfalls notwendige Kontexteinheit und Auswertungseinheit) und der Schritt der Hauptkategorienbildung übergangen. Der Verdacht kam bei mir auf, dass zur Erstellung der Vorlage ChatGPT benutzt worden war, und dann würde die Katze sich tatsächlich in den eigenen Schwanz beißen. Als jemand, der Jahrzehnte lang Einführungskurse in wissenschaftliches Arbeiten und Kolloquien zur Betreuung von Abschlussarbeiten an der Universität durchgeführt hat, möchte ich den vielen Nutzerinnen und Nutzern dieser Vorlage zurufen: Texte, die wissenschaftlichen Anspruch haben, müssen ausreichend mit Quellen versehen werden, die überprüfbar sind ("Mayring Mentor" enthält keine Literaturangaben!). Die Qualität der Quellen muss gesichert sein, wofür die Autor*innenschaft geprüft (Fachstudium, ev. Promotion, Habilitation, Lehrbefugnis) und die Qualität des Veröffentlichungsmediums erwogen werden sollte (Handbuch- oder Lehrbuchartikel, Fachzeitschrift mit Reviewverfahren, angesehener Fachverlag). Vor diesem Hintergrund ist "Mayring Mentor" nicht brauchbar. Die Tatsache, dass die Vorlage viele Nutzerinnen und Nutzer erreicht und gut bewertet wird, kann kein Kriterium sein. Dass ein reißerisches Boulevardblatt viele zufriedene Leserinnen und Leser erreicht, garantiert ja auch nicht den Wahrheitsgehalt der Artikel. [19]

5. Kodierübereinstimmung

Nicht ganz klar bin ich mir, ob durch einen erneuten Durchlauf der identischen Eingabe bei ChatGPT eher die Intracoder-Übereinstimmung (die gleiche Person analysiert das Material ein zweites Mal im Sinne einer Reliabilitätsprüfung) oder die Intercoder-Übereinstimmung (eine zweite Person analysiert das Material im Sinne einer Objektivitätsprüfung) festgestellt würde. In jedem Falle stellen beide wichtige Gütekriterien für qualitative Inhaltsanalyse dar (MAYRING 2022a, 2022b). Das Ergebnis des identischen Analyseauftrages einen Tag später ergab folgenden ChatGPT-Ausdruck (ChatGPT 3.5 vom 25. März 2024):

1. Vergleich mit Kollegen im Dienst
2. Gefühl von Wert und Kompetenz
3. Willkürliche Entlassung
4. Rechtliche Auseinandersetzung vor dem Arbeitsgericht
5. Unzufriedenheit mit aktuellen Tätigkeiten (nicht als Lehrer arbeiten)
6. Mangelnde Erfüllung durch alltägliche Aufgaben (Haushalt, Gartenarbeit)
7. Nutzung von Unterstützungsangeboten (Gewerkschaft, Weiterbildungsmaßnahmen)

² Ich bedanke mich für den Hinweis einer anonymen begutachtenden Person dieses Artikels.

8. Herausforderungen durch den Übergang von DDR zu Bundesrepublik im Arbeitskontext
9. Unsicherheit bezüglich des Arbeitsgerichtsverfahrens
10. Sorge um die finanzielle Situation der Familie [20]

Das Ergebnis war noch weniger brauchbar, da nicht nur Belastungsfaktoren, wie in der Anweisung gefordert, aufgezählt wurden. Manches wurde jetzt gefunden, was am Vortag übersehen worden war. Insgesamt muss man den Kodierübereinstimmungstest als gescheitert ansehen. Interessant ist, dass auch bei weiteren Überprüfungen der Kodierübereinstimmung auch mit Version 4 bei identischer Eingabe immer andere Lösungen angeboten wurden. Ich schließe daraus, dass ChatGPT höchstens grobe Annäherungen liefern kann. Ich schließe aus dem obigen Testlauf aber auch, dass ChatGPT Kodieranweisungen nicht immer befolgt und daraus grobe Fehler entstehen. Vielleicht lag dies aber auch an der frei verfügbaren Version 3.5, die ich verwendet hatte. [21]

6. Größere Materialmengen mit der kostenpflichtigen Version ChatGPT 4

ChatGPT zeigte, wie oben angeführt, eine Begrenzung der verarbeitbaren Materialmenge auf (4096 *Token*, etwa drei Seiten Transkript, wie mir rückgemeldet wurde). Mit der Bezahlversion 4 können immerhin ca. 500 Seiten Transkript bearbeitet werden. Es wird aber auf Anfrage darauf hingewiesen, dass die effektivste und genaueste Analyse nur bei kleineren Mengen gewährleistet sei und die Aufteilung des Materials in Portionen empfohlen. Für übergreifende Analysen ist das jedoch wenig hilfreich. Mit qualitativer Inhaltsanalyse werden oft mehr als tausend Seiten Material queranalysiert. [22]

Ich versuchte eine inhaltsanalytische Auswertung mit ca. 10 Seiten Interviewtranskripten, Ausschnitten von sechs Interviews mit arbeitslosen Personen über ihre Belastung. Hier stellte ich ChatGPT 4 die Frage nach einer Liste mit Belastungsfaktoren über die sechs Fälle hinweg, verbunden mit der Bitte, die Häufigkeit der gefundenen Kategorien aufzuführen, ein Standardverfahren induktiver Kategorienentwicklung. Hier die Antwort (ChatGPT 4, 18. März 2024):

- "1. Vergleich mit Kollegen, die noch im Dienst sind - 1 Mal
2. Willkürliche Entlassungen 1 Mal
3. Rechtliche Auseinandersetzungen - 1 Mal
4. Mangelnde Anerkennung der beruflichen Kompetenzen - 1 Mal
5. Schwierigkeiten bei der Jobsuche - 1 Mal
6. Unsicherheit bezüglich der Zukunft - 1 Mal
7. Finanzielle Unsicherheit - 1 Mal
8. Verlust des beruflichen Selbstbilds - 1 Mal
9. Fehlende berufliche Perspektiven - 1 Mal
10. Soziale Isolation - 1 Mal
11. Anpassung an neue berufliche Anforderungen - 1 Mal

12. Mangel an beruflicher Zufriedenheit - 1 Mal
13. Untätigkeit und Langeweile - 1 Mal
14. Druck, sich beruflich neu zu orientieren - 1 Mal
15. Familienbelastungen durch Arbeitslosigkeit - 1 Mal
16. Schwierigkeiten, mit dem Status der Arbeitslosigkeit umzugehen - 1 Mal
17. Erfahrung von Ungerechtigkeit und Entwertung früherer Tätigkeiten - 1 Mal
18. Notwendigkeit der Weiterbildung ohne konkrete Jobperspektive - 1 Mal
19. Belastung durch veränderte Lebensroutinen - 1 Mal
20. Verlust des beruflichen Netzwerks - 1 Mal" [23]

Es wurde eine Liste mit 20 induktiven Kategorien ausgegeben, jede entgegen der Anweisung jedoch nur einmal kodiert. Dies wurde so kommentiert: "Jeder dieser Punkte wurde als einmalig pro relevanter Erwähnung gezählt, was die Vielfalt der Belastungsfaktoren durch Arbeitslosigkeit bei den interviewten arbeitslosen Lehrern aufzeigt" (ChatGPT 4 vom 18. März 2024). Diese Auswertung stimmt mit der oben angeführten Musterlösung nicht überein. Fünf der 20 genannten Belastungsfaktoren (dort waren 16 Kategorien gefunden worden, s. Abschnitt 2) waren ähnlich, in der Musterlösung aber mehrfach in den Interviews genannt worden. So wurde die Textstelle³ "nur die Wohnung oder den Garten in Ordnung zu bringen füllt mich nicht aus" in Fall 1 nicht gefunden, in Fall 2 als "Untätigkeit und Langeweile" kodiert (Fall 2: "Ich könnte jetzt in den Garten gehen und Unkraut jäten, aber innerlich geht das nicht auf"). Vier der sechs Interviewten berichteten, dass sie ihre Kündigung als ungerecht ansahen, was ChatGPT 4 nicht erkannte. Viele Textstellen wurden nicht als Belastungsfaktoren ausgewertet, auch so eindeutige wie in Fall 1: "F: Hat die Arbeitslosigkeit auch positive Seiten für Sie? A: Eigentlich würde ich das verneinen, weil ich sehr sehr gerne Lehrer war, sehr sehr gerne unterrichtet habe". In der Musterlösung wurde das kodiert mit der Kategorie: "Verlust des geliebten Lehrerberufes". Dafür wurden Kategorien genannt, denen keine Textstellen eindeutig zugeordnet werden können. Letztlich ist die Auswertung von ChatGPT 4.0 wenig brauchbar. [24]

7. Fazit

ChatGPT hat in beiden Versionen (3.5 und 4) Ergebnisse produziert, die zum Teil als grobe Annäherungen, zum Teil aber auch als grobe Fehler einzustufen sind. Die größte Annäherung an die Musterlösung (mit "natürlicher Intelligenz") ergaben sich in der Version ChatGPT 3.5 bei kurzem Text und einfacher Fragestellung. Dass es nur Annäherungen sind zeigt sich auch darin, dass eine Reanalyse mit identischer Fragestellung bei ChatGPT in beiden Versionen unterschiedliche Ergebnisse zeitigte. [25]

Es muss dazu gesagt werden, dass die Möglichkeiten des Arbeitens mit ChatGPT bei weitem nicht ausgereizt wurden. Vor allem das Eingrenzen mit immer spezifischeren Fragen und die genaue *Prompt*-Gestaltung wurden wenig umgesetzt. Allerdings erfordert ein solches Vorgehen erhebliche Kompetenzen und genaueste Vorstellungen der Anwender*innen, was in der Regel selten

³ Die in diesem Abschnitt erwähnten Textstellen stammen aus meinen Projektunterlagen und wurden nicht veröffentlicht.

vorkommen dürfte. Insofern kann man von Fallstricken in der Verwendung von ChatGPT sprechen. Auch ein iteratives Annähern an die Ergebnisse entspricht nicht dem Forschungsstil der von unserer Arbeitsgruppe entwickelten qualitativen Inhaltsanalyse. Würde ein solches tentatives Verfahren angewendet, würde ich tendenziell nicht von einer qualitativen Inhaltsanalyse sprechen. [26]

Bei der komplexeren Fragestellung (induktive Kategorien zu Belastungsfaktoren durch Arbeitslosigkeit über mehrere Interviews hinweg) scheiterte auch ChatGPT 4. Wichtige Textstellen wurden übergangen, Kategorisierungen waren ungenau, Gemeinsamkeiten der Belastungsfaktoren zwischen den Interviews wurden (trotz Aufforderung) nicht gesehen. Speziellere Kodieranweisungen wurden nicht beachtet, verschiedene Ansätze der Textanalyse wurden nicht richtig angewendet. ChatGPT führte auch fehlerhafte Anweisungen durch (siehe oben "Analysiere bitte mit Themenanalyse nach Brown und Harris") und präsentierte Ergebnisse. [27]

Es sei auch auf die problematische Datenschutzsituation hingewiesen. Einerseits erfordert das Einsenden eines Interviewtranskriptes an ChatGPT eigentlich ein explizites, schriftliches Einverständnis der interviewten Person. Dass in dem in Abschnitt 2 erwähnten Fall ein 30 Jahre zurückliegendes Interview verwendet wurde, der Interviewte damals sein Einverständnis zur anonymisierten Verwendung des Interviews gegeben hatte und auch mittlerweile verstorben sein dürfte, täuscht nicht darüber hinweg, dass eine der [Datenschutz-Grundverordnung](#) konforme Nutzung von Interviewdaten in ChatGPT äußerst fraglich ist. Dort, wo ChatGPT inhaltliche Angaben über das Transkript hinaus macht ("Welche Schritte der Inhaltsanalyse nach Mayring?") stellte sich auch das Problem, dass die Antworten keine Literaturangaben vorwiesen und damit auch wissenschaftlich nicht verwertbar waren. [28]

Was kann ChatGPT? Nach den verschiedenen Versuchen erscheint es mir, dass grobe inhaltliche Kategorisierungen von Textstellen einigermaßen gelingen und so ein Überblick über Thematiken im Material gut gegeben werden kann. Das mag hilfreich sein, wenn man eine Auswertungstechnik wählt, die keine spezifischen Einschränkungen voraussetzt und eher im Deskriptiven bleibt. Dies scheint mir der Fall zu sein bei der thematischen Strukturierung nach KUCKARTZ (2014) oder der thematische Analyse nach BRAUN und CLARKE (2006). Wenn die einschlägigen Softwareentwickler und -entwicklerinnen z.B. bei ATLAS.ti und MAXQDA jetzt auf ChatGPT basierende KI-Module eingebaut haben, unterstützen sie damit nur eine bestimmte Art der Textanalyse. Der Einbau von Ablaufmodellen und Auswertungsregeln ist dort weniger vorgesehen. Das ist auch der Grund, weshalb für die von mir und meinem Team entwickelte qualitative Inhaltsanalyse ein eigenes, kostenlos zugängliches [Softwarepaket](#) entwickelt wurde, in dem die Ablaufschritte interaktiv bindend verankert sind. [29]

Wenig geeignet ist ChatGPT also, wenn

- spezifischere Techniken mit genaueren Auswertungsanweisungen wie beispielsweise in der qualitativen Inhaltsanalyse nach MAYRING (2022a) eingesetzt werden sollen;
- die Textanalyse mehr in die Tiefe gehen soll, auch unter Beachtung theoretischer Erwägungen, wie beispielsweise in der *Reflexive Theme Analysis* (BRAUN & CLARKE 2024), der GTM oder der psychoanalytischen Textinterpretation (vgl. MAYRING 2019b für die unterschiedlichen Ergebnisse verschiedener Textanalyseansätze, aufgezeigt an einem Beispieltext). Das strukturierte, theoriegeleitete, regelgeleitete Textanalyseverfahren erscheint mir aber zentral, da es hohe Wissenschaftsstandards (z.B. Kodierübereinstimmungstestung) erfüllen und sinnvoll in heute immer wichtiger werdende Mixed-Methods-Designs eingebaut werden kann. [30]

In der universitären Veranstaltung, die ich in Abschnitt 1 erwähnt habe, wurde das Argument eingebracht, man könne ja ChatGPT den Text erst einmal analysieren lassen und dann nacharbeiten. Dies erscheint mir jedoch wenig sinnvoll, da man ja nicht genau weiß, wo die Ungenauigkeiten und Fehler liegen, man also in jedem Fall das ganze Material durcharbeiten muss. Wozu dann der ChatGPT-Lauf? Eine grobe Annäherung an die Ergebnisse, kann, wenn in jedem Fall das gesamte Material bearbeitet werden muss, sogar einen verzerrenden Einfluss auf die Auswertung haben. [31]

Man könnte ChatGPT in der qualitativen Textanalyse für einen ersten, ganz offenen explorativen Schritt einsetzen. Die Anwendungsfälle für eine solche "Neulandforschung" sind jedoch in der Sozialforschung eher selten; meist geht man doch mit einer dezidierten Fragestellung und einem differenzierten Interviewleitfaden an das Material, und da wird ChatGPT ungenauer. Auch eine Theoriegeleitetheit der Analyse ist schwer umsetzbar. [32]

Ein letztes Argument könne die Menge des Materials sein, die mit ChatGPT verarbeitet werden kann. Zum einen ist dies nur in der Bezahlversion 4 möglich, und auch dort sind Einschränkungen festzustellen. Zum anderen kann es kein Ziel in der Sozialforschung sein, möglichst viel Material zu erheben. Für eine Wahlanalyse in der Bundesrepublik Deutschland werde ich nicht 50 Millionen Personen befragen, sondern eine Stichprobe auswählen. Auch in der qualitativen Forschung geht es darum, Stichproben so zu definieren, dass sie Verallgemeinerungen zulassen und eine tiefgehende, saubere Analyse zulassen. [33]

Abschließend sei betont, dass Werkzeuge wie ChatGPT laufend Überarbeitungen erfahren, weshalb dies nur eine zwischenzeitliche Bilanz sein kann. Ich denke aber, dass solche Erfahrungsberichte wichtig sind für die Einschätzung der möglichen Rolle von KI in der Wissenschaft. [34]

Literatur

- Braun, Virginia & Clarke, Victoria (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3, 77-101.
- Braun, Virginia & Clarke, Victoria (2024). Thematic analysis. In [Norman K. Denzin](#), Yvonna S. Lincoln, Michael D. Giardina & Gaile S. Canella (Eds.), *The Sage handbook of qualitative research* (6. Aufl., S.385-402). Los Angeles, CA: Sage.
- Brown, George W. & Harris, Tirril (1978). *Social origins of depression: A study of psychiatric disorder in women*. New York, NY: Free Press.
- Glaser, Barney G. & Strauss, Anselm L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Chicago, IL: Aldine.
- Haubl, Rolf & Lohl, Jan (2020). Tiefenhermeneutik. In [Günter Mey](#) & [Katja Mruck](#) (Hrsg.), *Handbuch Qualitative Forschung in der Psychologie* (2. Aufl., Bd. 2, S.555-578). Wiesbaden: Springer VS.
- Kuckartz, Udo (2012). *Qualitative Inhaltsanalyse. Methoden, Praxis, Computerunterstützung*. Weinheim: Beltz Juventa.
- Kuckartz, Udo (2014). *Qualitative text analysis. A guide to methods, practice & using software*. London: Sage.
- Lieder, Fabio R. & Schäffer, Burkhard (2024). *Reconstructive social research prompting (RSRP). Distributed interpretation between AI and researchers in qualitative research*, <https://doi.org/10.31235/osf.io/d6e9m> [Zugriff: 1. Oktober 2024].
- [Mayring, Philipp](#) (2019a). Qualitative Inhaltsanalyse – Abgrenzungen, Spielarten, Weiterentwicklungen. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, 20(3), Art. 16, <https://doi.org/10.17169/fqs-20.3.3343> [Zugriff: 1. Oktober 2024].
- Mayring, Philipp (2019b). Zentrale qualitative Auswertungsverfahren. In Marius Harring, Carsten Rohlf & Michaela Gläser-Zikuda (Hrsg.), *Handbuch Schulpädagogik* (S.859-868). Münster: Waxmann – utb.
- Mayring, Philipp (2022a). *Qualitative content analysis. A step-by-step guide*. London: Sage.
- Mayring, Philipp (2022b). *Qualitative Inhaltsanalyse. Grundlagen und Techniken* (13., überarbeitete Auflage). Weinheim: Beltz.
- Mayring, Philipp & Fenzl, Thomas (2022). Qualitative Inhaltsanalyse. In Nina Baur & Jürgen Blasius (Hrsg.), *Handbuch Methoden der empirischen Sozialforschung* (3. vollst. überarb. u. erw. Aufl., S.691-706). Wiesbaden: Springer VS.
- Mayring, Philipp; König, Joachim; Birk, Nils & Hurst, Alfred (2000). *Opfer der Einheit – Eine Studie zur Lehrerarbeitslosigkeit in den neuen Bundesländern*. Opladen: Leske + Budrich.
- [Schreier, Margrit](#) (2012). *Qualitative content analysis in practice*. London: Sage
- Strauss, Anselm (1991 [1987]). *Grundlagen qualitativer Sozialforschung. Datenanalyse und Theoriebildung in der empirischen soziologischen Forschung*. München: Fink.

Zum Autor

[Philipp MAYRING](#) ist Professor für psychologische Forschungsmethoden im Ruhestand am Institut für Psychologie der Alpen-Adria Universität Klagenfurt sowie Leiter des gemeinnützigen Vereins zur Förderung qualitativer Forschung/Association for Supporting Qualitative Research (ASQ) in Klagenfurt. Seine Forschungsschwerpunkte liegen in den Bereichen qualitative Textanalyse, Evaluationsforschung, Gesundheits- sowie Wohlbefindensforschung.

Kontakt:

Alpen-Adria Universität Klagenfurt
Institut für Psychologie
Universitätsstr. 65-57, 9020 Klagenfurt,
Österreich

E-Mail: philipp.mayring@aau.at

URL: <https://philipp.mayring.at/>

FQS 26(1), Art. 4, Philipp Mayring: Qualitative Inhaltsanalyse mit ChatGPT: Fallstricke, grobe Annäherungen und grobe Fehler. Ein Erfahrungsbericht

Zitation

Mayring, Philipp (2025). Qualitative Inhaltsanalyse mit ChatGPT: Fallstricke, grobe Annäherungen und grobe Fehler. Ein Erfahrungsbericht [34 Absätze]. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, 26(1), Art. 4, <https://doi.org/10.17169/fqs-26.1.4252>.