VORTRAGSSKRIPT

KI-Nutzungsstrategien in qualitativer Forschung

Angebote, Herausforderungen und methodische Nutzungsweisen

2. NOVEMBER 2026

DR. THORSTEN DRESING

https://www.audiotranskription.de/ki-einsatz-in-der-qualitativen-forschung/

Inhalt

Vorwort 3

Worum es in diesem Vortragsskript geht 3

1 **KI-Nutzungsangebote & Studien zur KI-Nutzung** 5

Von der automatischen Transkription zur automatisierten Auswertung 5

Zuerst die Rahmenbedingungen des Datenschutzes 6

Welche Angebote gibt es bereits 9

Eine qualitative Evaluation mit einem lokalen Modell bearbeiten 10

Ein integriertes Werkzeug am Beispiel von MAXQDA Tailwind 12

Fünf Studien zur Nutzung von KI 14

2 **Herausforderungen beim Einsatz von KI** 18

Eine Auswahl der Herausforderungen 18

A - Menge und Tempo der Entwicklung 21

B - Das Promptproblem am Beispiel einer Zusammenfassung 22

C - Das Prüfproblem von LLM-Ausgaben 25

3 **Aktuelle Ansätze** 30

Methodisierte Integrations- und Innovationsvorschläge 30

Was die Vorschläge leisten und was wir mitnehmen können 37

Anhang 39

Quellen und Materialien zum Nacharbeiten 39

Vorwort

Dieses Skript wurde mit KI-Unterstützung erstellt. Ich habe meinen Vortrag vom 2. November 2026 an der Universität Rostock aufgenommen, automatisch mit f4 transkribiert, das Transkript durch Claude (5.1), ChatGPT (6.1) und Gemini 4 Argon überarbeiten lassen, Links ergänzt und einmal manuell redigiert. Es ist die vierte wesentlich überarbeitete Fassung von Skripten zu diesem Thema und hält meinen thematischen Stand vom November 2026 fest. Diese und ältere Fassungen stehen unter https://www.audiotranskription.de/ki-einsatz-in-der-qualitativen-forschung/ zur freien Verfügung. Es geht inhaltlich um KI-Angebote für die qualitative Datenanalyse, die damit verbundenen Herausforderungen und Vorschläge für einen methodisch begründbaren Einsatz. Ich möchte informieren, zum eigenen Ausprobieren anregen und auch irritieren.

Der Gedankengang, die Beispiele und die persönliche Ansprache des Vortrags sind erhalten geblieben. Wiederholungen und organisatorische Passagen wurden gestrafft, einzelne Formulierungen präzisiert und Quellen mit Links ergänzt. Das konkrete Serendipity-Beispiel aus den Folien ist im dritten Teil vollständig abgedruckt; die detaillierte Aufschlüsselung der Aufwandsschätzung und die abschließenden Orientierungspunkte finden sich auf den Vortragsfolien. Die Mündlichkeit bleibt an manchen Stellen bewusst stehen.

Warum erzähle ich überhaupt etwas zu KI und qualitativer Forschung? Ich habe in Marburg studiert (2000) und promoviert (2006) und war dort wissenschaftlicher Mitarbeiter bei Udo Kuckartz. Seitdem ist ein Teil meiner Berufsbiografie die Auseinandersetzung mit qualitativer Forschung, qualitativer Inhaltsanalyse und Programmen, die wir dafür einsetzen. Einige kennen mich auch aus den MAXQDA-Workshops, die ich regelmäßig gebe. Der zweite Teil meiner Arbeit dreht sich um die Transkription. Früher um das manuelle Transkribieren mit unterschiedlichen Regelsystemen, heute auch um automatische Transkription und die technischen Lösungen, die wir bei audiotranskription entwickeln. Seit einigen Jahren ist KI darin ein fester und sinnvoller Bestandteil. In den letzten zweieinhalb Jahren ist KI für mich im Bereich qualitativer Auswertung im Blick und seit Anfang 2025 stelle ich meine Eindrücke dazu in Vorträgen dar. Die Inhalte der Vorträge entwickeln sich so wie die Technik und Methodik weiter, neue Funktionen kommen hinzu, neue methodische Vorschläge werden veröffentlicht und manches stellt sich schon wenige Wochen später anders dar. Deswegen lasse ich immer wieder einen Vortrag transkribieren, redigiere den Text und stelle ein Skript zur Verfügung. Auch die Quellen sollen zugänglich sein, damit Sie einzelne Aspekte vertiefen und die Studien selbst nachlesen können.

Ich bin Mitgründer von audiotranskription und an der Entwicklung von f4 beteiligt. Auch am hybriden Interpretieren, am Serendipity-Prompting und an einigen der besprochenen Veröffentlichungen habe ich mitgearbeitet. Dies alles bedingt eine spezifische Position.

Worum es in diesem Vortragsskript geht

Ich möchte das Thema möglichst differenziert und beispielbasiert vorstellen. Es geht darum, was Menschen heute schon mit KI machen können, welche Herausforderungen damit verbunden sind und wie die wissenschaftliche Community versucht, damit umzugehen.

Im ersten Teil schauen wir uns zunächst die Rahmenbedingungen an, insbesondere den Datenschutz. Anschließend zeige ich, was bereits heute möglich ist. Was könnte ein Mensch

heute machen, der sagt: „Ich möchte jetzt KI einsetzen“? Wir schauen uns solche Funktionen an und vergleichen die Geschwindigkeit mit dem Aufwand, den entsprechende Arbeiten bisher verursacht haben. Ich rahme dann diese exemplarischen Beispiele durch fünf Studien aus den Jahren 2025 und 2026, die auf unterschiedliche Weise Auskunft darüber geben, was der Einsatz von KI eigentlich mit uns macht.

Im zweiten Teil kommen die Herausforderungen. Ich gebe zunächst einen Überblick und greife dann drei Felder heraus: die Menge und das Tempo der Entwicklung, die Gestaltungsherausforderungen der Promptentwicklung und die Prüfung der LLM-Ausgaben. Gerade die beiden letzten Punkte sind für qualitative Forschung besonders zentral.

Im dritten Teil geht es um methodisierte Integrations- und Innovationsvorschläge in bestehende Verfahren qualitativer Datenanalyse. Wir haben seit vielen Jahrzehnten eine Diskussion über Gütekriterien und Verfahrensweisen qualitativer Forschung. Es gibt unterschiedliche methodische Ansätze, und die Wahl eines gegenstandsangemessenen Verfahrens bestimmt viele weitere Entscheidungen. Wenn ich nun ein Interview bei einem Sprachmodell hochlade, ist damit noch keine qualitative Forschungsmethodik gewählt. Ich muss eine Verbindung herstellen. Was ist meine Methode, und welche Aufgabe soll KI darin übernehmen? In der Literatur sehen wir dazu Integrationsvorschläge, die KI in bestehende Verfahren einbeziehen, und Innovationsvorschläge, die neue Formen der Zusammenarbeit entwickeln. Zwei der Vorschläge, die ich später vorstelle, haben wir selbst mitentwickelt: das hybride Interpretieren mit drei dialogisch integrierten Large Language Models und das Serendipity-Prompting.

Spätestens am Ende sollte deutlich sein, dass wir keine fertige allgemeine Lösung haben. Wir müssen Entscheidungen treffen, für die es innerhalb der wissenschaftlichen Community noch keinen breiten Konsens gibt. Wir befinden uns mitten in einer Übergangsphase, in der sich etablierte Forschungspraktiken verändern und wir gut daran tun, diese Veränderungen mitzugestalten.

1

TEIL 1

KI-Nutzungsangebote & Studien zur KI-Nutzung

Von der automatischen Transkription zur automatisierten Auswertung
Zuerst die Rahmenbedingungen des Datenschutzes
Welche Angebote gibt es bereits
Eine qualitative Evaluation mit einem lokalen Modell bearbeiten
Ein integriertes Werkzeug am Beispiel von MAXQDA Tailwind
Fünf Studien zur Nutzung von KI

Von der automatischen Transkription zur automatisierten Auswertung

Eine dieser Veränderungen kennen Sie vermutlich alle: die automatische Transkription. Über viele Jahrzehnte haben wir Interviews manuell transkribiert, also mit den eigenen Fingern abgetippt und dabei bestimmte Regelsysteme angewendet. Das war ein zentraler Arbeitsschritt qualitativer Forschung, und er konnte je nach Regelsystem ziemlich lange dauern.

Heute sehen wir, dass sehr viele Menschen zunächst automatisch transkribieren lassen und anschließend einen Korrekturdurchgang machen. Dabei überprüfen sie die Ausgabe an der Aufnahme und wenden das Regelsystem an, das für die eigene Auswertungsmethodik passend ist. Am Ende soll das Material so vorliegen, dass wir damit weiterarbeiten können.

Die automatische Transkription ist nicht perfekt. Wir haben aber einen Umgang mit ihren Fehlern entwickelt. Wir hören die Aufnahme und korrigieren den Text. Mit diesem Verfahren lässt sich gegenüber dem vollständigen manuellen Abtippen erheblich Zeit sparen. Als Erfahrungswert aus unserer Arbeit nenne ich häufig ungefähr die Hälfte. Das ist keine repräsentative weltweite Erhebung, sondern eine Größenordnung aus der Praxis. Je nach Aufnahmequalität und Transkriptionsregeln kann der Gewinn anders ausfallen.

Ich biete seit 22 Jahren Lösungen für die Transkription an, früher für das manuelle, heute auch für das automatische Transkribieren. Aus unserer Arbeit weiß ich, dass über uns bereits mehr als 100.000 Menschen aus dem Hochschulkontext automatische Transkription

eingesetzt haben. Auch in den Workshops begegnet mir der Einsatz ständig. Wenn mit Interviews gearbeitet wird, ist irgendeine Form automatischer Transkription fast immer mindestens in der Erwägung. Die genaue Verbreitung kennen wir damit noch nicht, aber es handelt sich um eine relevante Veränderung der Arbeitspraxis.

Im Vergleich zu anderen KI-Anwendungen ist die Akzeptanz hoch. Es gibt selbstverständlich Fragen zum Datenschutz und zur Qualität. In vielen Fällen wird der Einsatz aber nicht grundsätzlich abgelehnt.

Da liegt der nächste Gedanke nahe, denn die Transkription ist ja gar nicht unbedingt der aufwendigste Teil qualitativer Forschung. Mehr Zeit verbringen wir mit der Auswertung, also der Auseinandersetzung mit dem Material. Wenn KI bei der Transkription schon so gut hilft, könnte sie uns dann auch bei der Auswertung Arbeit abnehmen?

Genau das schauen wir uns an. Bei der Transkription haben wir eine Aufnahme, mit der wir die Wiedergabe vergleichen können. Bei einer Auswertung kommt etwas hinzu: Wir müssen beurteilen, welche Inhalte ausgewählt, miteinander verknüpft und wie sie gedeutet werden. Ob das genauso gut kontrollierbar ist, wird uns noch beschäftigen.

Zuerst die Rahmenbedingungen des Datenschutzes

Ich beginne mit einem Punkt, den viele als unangenehm empfinden: dem Datenschutz. Dabei möchte ich vor allem mit zwei oder drei verbreiteten Vereinfachungen aufräumen.

Wenn wir KI einsetzen, läuft das Modell häufig nicht auf unserem eigenen Rechner. Vielleicht stellt die Hochschule eine Infrastruktur bereit, etwa über ihr Rechenzentrum. Vielleicht nutzen wir Angebote wie die Academic Cloud der GWDG. Oder wir verwenden Dienste großer Anbieter wie ChatGPT, Claude oder Gemini. Auch Transkriptions- und QDA-Programme können externe oder hochschulinterne KI-Dienste ansprechen.

In diesen Fällen müssen wir uns klarmachen, dass hier in der Regel eine Datenübertragung stattfindet. Wohin werden die Daten übertragen, wer verarbeitet sie, und auf welcher Grundlage geschieht das? Bei qualitativen Interviews haben wir es in der Regel mit personenbezogenen Daten zu tun. Zwar gilt die DSGVO auch bei lokaler Verarbeitung, aber bei der Nutzung eines externen Dienstes kommen Fragen zur Übermittlung und zum Dienstleister hinzu.

Mindestens zwei Ebenen sollten wir unterscheiden. Die erste betrifft das technische und organisatorische Angebot. Wie sind die Daten geschützt? Wer hat Zugriff? Wo werden sie verarbeitet? Werden sie gespeichert, wann werden sie gelöscht und werden sie zum Training verwendet? Welche vertraglichen Regelungen gelten, und wird gegebenenfalls ein Auftragsverarbeitungsvertrag benötigt? Serverstandorte oder Zertifizierungen können dabei wichtige Anhaltspunkte sein. Ein technisch abgesicherter Dienst, der sich DSGVO-konform nennen darf, bedeutet aber noch nicht, dass wir unsere Interviews dort auch verarbeiten dürfen.

Bei qualitativen Interviews ist zusätzlich eine informierte Einwilligung der Regelfall. Wenn unsere Verarbeitung auf Einwilligung beruht, muss sie die geplante KI-Verarbeitung abdecken. Andere gesetzliche Rechtsgrundlagen kommen in bestimmten Konstellationen

infrage und müssen konkret geprüft werden. Ein Auftragsverarbeitungsvertrag allein ersetzt diese Prüfung nicht.

Informiert heißt eben nicht nur: „Kannst du bitte diese Seite unterschreiben, damit ich deine Daten verwenden darf?“ Wir müssen verständlich beschreiben, was wir mit welchen Daten tun wollen und welche Dienste daran beteiligt sind. Zum Beispiel: Ich möchte einen bestimmten KI-Dienst einsetzen, um fallbezogene Zusammenfassungen zu erstellen oder ein vorhandenes Codesystem auf das Material anzuwenden. Was wird dafür übermittelt? Zu welchem Zweck? Wie lange werden die Daten gespeichert? Wann werden sie gelöscht? u. v. m.

Außerdem müssen wir uns mit möglichen Folgen für die Befragten auseinandersetzen. Könnte eine Verarbeitung Nachteile haben? Was wäre, wenn Informationen entgegen der Planung zugänglich würden? Und was bedeutet das bei sensiblen Themen? Je nach Vorhaben können weitergehende Anforderungen bis hin zu einer Datenschutz-Folgenabschätzung hinzukommen.

Ich würde das nicht im Alleingang entscheiden. Vorhaben, eingesetzte Dienste, Speicher- und Löschfristen und die erforderlichen Verträge sollten wir mit den Datenschutzbeauftragten abstimmen. Gegebenenfalls gehört auch die Ethikkommission dazu. Ein Ethikvotum ersetzt dabei nicht die datenschutzrechtlichen Anforderungen, kann aber helfen, die Forschungssituation und ihre möglichen Folgen zu beurteilen.

Das ist bereits eine ganze Menge Arbeit. Und deshalb gibt es eine zunächst ausgesprochen plausible Idee: Dann anonymisiere ich einfach die Daten. Ich entferne Namen, Orte und Firmen, und danach sind die Inhalte doch nicht mehr personenbezogen.

Warum das Entfernen von Namen und Orten nicht genügt

Genau an dieser Stelle wird es bei qualitativen Interviews schwierig. Eine Stunde Interview ergibt je nach Darstellung ungefähr 15 bis 20 Seiten Text. Auf diesen Seiten stehen nicht nur offensichtliche Identifikatoren wie Name und Wohnort. Es stehen auch biografische Ereignisse, berufliche Tätigkeiten, zeitliche Abläufe, besondere Erfahrungen und Beziehungen zu anderen Personen darin. Eine Kombination solcher Einzelheiten kann zur Reidentifikation führen, selbst wenn wir für eine Anonymisierung offensichtliche Identifikationsmerkmale entfernt haben.

Ich nehme zunächst ein bewusst einfaches Beispiel. Wir lesen einen anonymisierten Interviewauszug über eine ehemals sehr aktive politische Person. Dort steht, dass sie ihre Hände in Form einer Raute hält und einer anderen Person ihr Vertrauen ausspricht. Vermutlich haben Sie sofort eine Assoziation: Angela Merkel. Name, Parteizugehörigkeit und Funktion mussten dafür gar nicht genannt werden. In diesem zugespitzten Beispiel ist die Geste bereits sehr charakteristisch. Es macht aber anschaulich, dass Erkennbarkeit nicht nur an Namen hängt.

Eine Studie hat das für einen Interviewdatensatz konkreter untersucht. Es geht um das öffentlich bereitgestellte Material des Anthropic-Interviewer-Projekts mit 1.250 Interviews. Die Untersuchung von Li fragt, ob agentische Sprachmodelle trotz der anonymisierten Transkripte herausfinden können, wer hinter einzelnen Interviews steht. Die Modelle verbinden dafür Hinweise aus den Texten mit weiteren Informationen. Untersucht wurde eine gezielt ausgewählte Teilmenge von 24 publikationsbezogenen Interviews aus 125

Interviews mit Wissenschaftlerinnen und Wissenschaftlern. In sechs Fällen konnten damit auch Personen identifiziert werden. Das ist kein Nachweis einer Reidentifikation aller Befragten, aber ein deutliches Warnsignal. (Li 2026)

Die Kombination scheinbar harmloser Einzelheiten kann also entscheidend sein. Eine Person erzählt etwa von einer Dreiergruppe an der Universität, einem besonderen Ereignis im Seminar und einer verspäteten Abgabe. Für sich genommen klingt keiner dieser Aspekte wie eine direkte Personenangabe. In der Verbindung können sie aber Menschen mit Kontextwissen helfen, eine konkrete Situation und damit Beteiligte zu erkennen.

Unser eigener Beitrag setzt sich mit genau diesem Problem auseinander. Namen und Orte zu maskieren kann das Risiko senken. Es ist aber keine ausreichende Grundlage, um bei detailreichen qualitativen Daten ohne weitere Prüfung von Anonymität auszugehen. (Dresing / Krähnke / Hense / Pehl 2026) (Dresing / Pehl / Krähnke 2026)

Dann könnte die nächste Idee lauten: Gut, dann anonymisieren wir eben noch gründlicher. Wir löschen oder verallgemeinern auch die Kontextinformationen. Das kann allerdings zu einem anderen Problem führen. Gerade qualitative Forschung braucht häufig diese Einzelheiten. Welche Wörter verwendet jemand? In welcher Reihenfolge wird argumentiert? Welche Situation wird beschrieben? Was unterscheidet einen Fall von einem anderen?

Wenn wir all das so weit verändern, dass es keine Rückschlüsse mehr ermöglichen soll, kann genau das Material verloren gehen, das wir für die Analyse brauchen. Ein Fallbericht von Laudel und Bielick macht das besonders anschaulich. Sie schickten ein anonymisiertes Testinterview an Qualiservice in Bremen und fragten, ob es so noch sozialwissenschaftlich nutzbar sei und ob sie die Anonymisierung auf diese Weise fortsetzen sollten. Die Rückmeldung lautete, dass das aufbereitete Interview für keine sozialwissenschaftliche Fragestellung mehr geeignet sei. Die Forschenden beendeten daraufhin ihren Archivierungsversuch. Das war ein konkreter Fall, aber er zeigt den Zielkonflikt sehr deutlich. (Laudel / Bielick 2019)

Mit diesem Zielkonflikt beschäftigt sich auch die Qualiservice-Handreichung zur Anonymisierung qualitativer Textdaten. Schutz und wissenschaftlicher Nachnutzungswert müssen zusammen betrachtet werden. Eine stärkere Bereinigung kann den Text so weit verändern, dass er für eine Sekundäranalyse kaum noch brauchbar ist. (Mozygemba / Hollstein 2023)

Ich würde daraus kein pauschales Urteil über jedes qualitative Material ableiten. Die praktische Warnung bleibt aber: „Dann anonymisiere ich einfach" ist bei längeren (Interview-)Texten keine hinreichende Antwort auf die Frage, ob wir sie einem KI-Dienst überlassen dürfen. Gerade bei sensiblen Themen kann eine belastbare Anonymisierung sehr schwierig sein, und manchmal können auch wenige Sätze ausreichen, um eine Identifikation zu ermöglichen.

Welche Wege bleiben

Welche Möglichkeiten haben wir dann? Eine Möglichkeit ist, die externe Verarbeitung auf einer passenden Grundlage abzusichern. Das kann je nach Konstellation eine hinreichend informierte Einwilligung oder eine andere einschlägige Rechtsgrundlage sein. Hinzu kommen die erforderlichen Schutzmaßnahmen und gegebenenfalls Regelungen zur Datenübermittlung. Es gibt also kein allgemeines Verbot jeder externen KI-Nutzung. Aber

die Aussage „Die Befragten haben unterschrieben“ klärt auch nicht automatisch alle Anforderungen. Wie realistisch eine solche Lösung ist, hängt vom Vorhaben ab. In qualitativer Forschung arbeiten wir häufig mit sensiblen Themen, etwa mit Gesundheit oder persönlichen psychologischen Erfahrungen. Und ob uns dabei die Befragten eine solche KI-Nutzung unterschreiben (sollten), halte ich für fragwürdig. Und selbst bei einem heute vermeintlich harmlosen Thema ist nicht sicher, ob bestimmte Informationen einer Person in zehn oder zwanzig Jahren Nachteile bringen könnten. Gesellschaftliche Situationen können sich ändern. Auch solche Überlegungen gehören zu einer verantwortbaren Einschätzung der Folgen.

Die zweite Möglichkeit ist, lokal zu arbeiten. Dann laden wir ein geeignetes Modell auf den eigenen Rechner und verarbeiten die Daten dort, ohne sie an einen externen KI-Dienst zu übermitteln. Das kann gerade bei nicht anonymisierten Interviews ein wichtiger Weg sein. Die datenschutzrechtlichen und institutionellen Voraussetzungen für die Verarbeitung bestehen weiterhin, die zusätzliche Übermittlung an den KI-Dienst entfällt aber.

Für größere lokale Modelle braucht man entsprechende Hardware. Ich arbeite in meinem Beispiel mit einem Rechner mit 128 Gigabyte Arbeitsspeicher. Das ist eine erhebliche Investition. Ich zeige gleich, wie man ein lokales Modell mit einer kostenfrei verfügbaren Oberfläche nutzen kann. Ob diese Ausstattung nötig ist, hängt vom Modell und der Aufgabe ab; kleinere Modelle kommen mit weniger Ressourcen aus.

Die dritte Möglichkeit betrifft die Auswahl des Materials. Zum Erproben müssen wir nicht mit sensiblen Originalinterviews anfangen. Wir können vollständig fiktive Interviews oder andere geeignete Texte ohne Personenbezug verwenden. Es ist auch möglich, fiktive Interviews durch Sprachmodelle erzeugen zu lassen und anschließend damit zu arbeiten. Als Experimentierfeld kann das interessant sein.

Damit haben wir den Rahmen. Ich will keine pauschale Verbotsbotschaft vermitteln. Ich will deutlich machen, dass wir die Bedingungen klären müssen, bevor wir Daten hochladen, und dass dies aufwendig ist. Ich erlebe in der Praxis aber nicht gerade selten, dass diese Auseinandersetzung unterbleibt und Daten ohne geeignete Grundlage übertragen werden.

Welche Angebote gibt es bereits

Schauen wir uns an, was heute bereits verfügbar ist. Automatische Transkription kennen Sie vermutlich. Aber was gibt es darüber hinaus für die Auswertung?

Es beginnt mit etwas scheinbar Banalem, einer Zusammenfassung. Viele haben schon einen Artikel, ein PDF-Dokument oder einen anderen Text zusammenfassen lassen. Zusätzlich können wir Paraphrasen erstellen lassen, Fragen an das Material stellen und Antworten darauf bekommen. Wir können Themen identifizieren und ausdifferenzieren lassen. Auch Belegstellen lassen sich filtern und zuordnen. Wenn ich bereits Codes mit Definitionen und Operationalisierungen habe, kann ein Werkzeug nach entsprechenden Stellen suchen und sie automatisch codieren.

Vielleicht folge ich einer eher rekonstruktiven Strategie, dann schaue ich sehr genau auf kleinere Textstellen und kann mit KI Lesarten generieren lassen oder eigene Lesarten kontrastieren und Interpretationsvorschläge ergänzen.

Und schließlich gibt es die One-Button-Lösungen. Ich gebe das Material hinein und lasse mir einen vollständigen Ergebnisbericht erzeugen.

Das sind keine bloßen Zukunftsideen. Solche Funktionen werden bereits angeboten und eingesetzt. Ob ihre Ausgaben für einen bestimmten Forschungsprozess angemessen sind, ist damit noch nicht beantwortet.

Direkte Arbeit mit einem Modell oder vorgefertigte Funktionen

Die Zugänge zu diesen Funktionen lassen sich für unseren Zweck in zwei Gruppen aufteilen. Erstens können Menschen ein Sprachmodell über eine Chatoberfläche ansprechen. Das kann ein Hochschulangebot, die Academic Cloud oder ein kommerzieller Dienst sein. Auch eine lokale Oberfläche kommt infrage. Wer technisch weiter einsteigt, arbeitet vielleicht über ein Terminalfenster. Andere nutzen bereits agentische Workflows und Loops, in denen mehrere Arbeitsschritte aufeinanderfolgend und iterativ verbessernd abgearbeitet werden. Dabei stellt sich die Frage, wie man Modelle auswählt, Prompts gestaltet, Ergebnisse beurteilt und den Ablauf so organisiert, dass brauchbare Ausgaben entstehen. Das kann technisch und methodisch aufwendig sein. Promptentwicklung, Hardware und Zugänge spielen eine Rolle. Verständlicherweise gibt es Menschen, die sagen, damit möchte ich mich nicht selbst beschäftigen. Dafür haben doch andere (Firmen oder Menschen) schon etwas entwickelt.

Damit kommen wir zum zweiten Zugang: dem über Softwareanwendungen mit vorbereiteten Funktionen. Bei diesen haben die EntwicklerInnen die LLMs ausgewählt, Prompts entwickelt und das Ganze in einer Oberfläche zugänglich gemacht. Dazu gehören etwa MAXQDA AI Assist und Tailwind, ATLAS.ti, f4, NVivo, Qinsights oder KarlAI. Die Angebote wachsen, um es anschaulich zu sagen, wie Giersch im Garten. Für Nutzende ist das bequem. Ich muss nicht zuerst selbst den Prompt für eine Zusammenfassung schreiben. Ich drücke auf einen Button. Damit sind die methodischen Entscheidungen aber nicht verschwunden. Sie wurden an anderer Stelle getroffen. Darauf kommen wir später zurück.

Ich zeige jetzt zwei Beispiele. Zuerst die lokale Arbeit mit einem selbst entwickelten Prompt, dann die Arbeit mit vorbereiteten Funktionen.

Eine qualitative Evaluation mit einem lokalen Modell bearbeiten

Für den lokalen Zugang kann man sich zum Beispiel LM Studio herunterladen. Die Oberfläche sieht für Menschen, die bereits mit Chatbots gearbeitet haben, vertraut aus. Unten gibt es ein Eingabefeld, in dem ich meine Anfrage formuliere. Über einen zusätzlichen Button kann ich Material hinzufügen. Zuvor wähle ich ein geeignetes Sprachmodell aus, das auf meinem Rechner ausgeführt wird. (LM Studio)

Eine große Plattform für solche Modelle ist Hugging Face. Dort werden viele Open-Weight-Modelle angeboten. Wir können ein Modell herunterladen und müssen dabei unter anderem auf Größe, technische Anforderungen und die jeweilige Lizenz achten. Wir probieren regelmäßig neue Modelle aus, um ihre Möglichkeiten für unsere Aufgaben kennenzulernen. Die Auswahl ist inzwischen groß, und sie entwickelt sich fortlaufend weiter. (Hugging Face)

Was habe ich in diesem Beispiel gemacht? Ich habe einen älteren eigenen Datensatz genommen. Udo Kuckartz, Stefan Rädiker, Claus Stefer und ich haben damals eine qualitative Lehrveranstaltungsevaluation durchgeführt und den Ablauf in einem Buch dokumentiert. Wir wollten einen möglichst nachvollziehbaren, deskriptiven Einstieg in qualitative Datenanalyse zeigen. Das Material haben wir ebenfalls zugänglich gemacht. (Kuckartz u. a. 2008)

Die Evaluation hatten wir damals manuell durchgeführt. Und wir haben dabei mit der Stoppuhr gemessen. Wie lange brauchen wir für die Entwicklung des Leitfadens, die Interviews, die Transkription, die Fallzusammenfassungen, das Codesystem, das Codieren, die Ausdifferenzierung und den Ergebnisbericht? Also die Schritte des gesamten empirischen Arbeitsteils.

Dabei kam ein Zeitfaktor von ungefähr eins zu sechzig heraus. Die Eins steht für eine Stunde Audiodaten, die Sechzig für die damit verbundenen Arbeitsstunden im untersuchten Ablauf. Das war schon damals eine Zahl, bei der einige Menschen blass geworden sind. Übertragen wir diese Größenordnung einmal auf zehn Stunden Interviewmaterial, also etwa zehn Interviews von jeweils einer Stunde oder zwanzig Interviews von je einer halben Stunde. Dann wären das ungefähr 600 Arbeitsstunden bis zum Ergebnisbericht. Das ist keine allgemeingültige Vorschrift für jedes qualitative Projekt. Andere methodische Zugänge und Forschungsfragen können ganz andere Aufwände mit sich bringen. Es zeigt aber, über welche Größenordnungen wir sprechen.

Wir konnten diese Arbeit auch nicht acht Stunden täglich mit derselben Konzentration erledigen. Wenn im Durchschnitt über einen längeren Zeitraum vier Stunden konzentrierte Arbeit pro Tag möglich sind, verteilen sich 600 Arbeitsstunden schnell über viele Monate. Das dürfte einigen bekannt vorkommen, die qualitativ forschen.

Meine Fantasie bei der erneuten Beschäftigung mit diesem Datensatz war ausgesprochen einfach. Ich möchte diese vielen Arbeitsstunden auf wenige Minuten reduzieren. Für die Reanalyse habe ich zehn Kurzinterviews mit ungefähr 35 Seiten Text genommen und einen längeren Prompt entwickelt. Darin stehen mehrere Arbeitsaufträge und der Prompt allein umfasst mehrere Seiten. Dahinter folgt das Interviewmaterial, nach Absätzen nummeriert. So kann das Modell einzelne Aussagen auf konkrete Stellen beziehen.

Die lokale Maschine verarbeitet nun diesen Prompt und das Material. In der Demonstration hat das Modell nach dem Thinking-Prozess eine längere Ausgabe in Form eines Evaluationsberichtes erzeugt. Es wurde zum Beispiel ein Themenbereich zur emotionalen Bewertung und zur Veränderung der Einstellung im Zeitverlauf entwickelt. Solche Themenbezeichnungen hatte ich nicht als fertiges Ergebnis vorgegeben. Die Ausgabe enthält Ausführungen dazu und Belegzitate aus dem Material.

Das geht lokal. Die Daten werden in dieser Konstellation nicht an einen externen KI-Dienst übertragen. Und die Ausgabe entsteht sehr schnell, wenn das Modell und der Rechner dafür geeignet sind.

Allerdings fällt ein solcher Prompt nicht einfach aus der Hand. An seiner Entwicklung habe ich über Monate gearbeitet und bin immer noch nicht zufrieden. Mit einem anderen Modell kann sich die Ausgabe wieder verändern. Das ist zum Teil eine Sisyphusarbeit. Man formuliert etwas, testet es, findet ein Problem, ändert den Prompt, und dann taucht an einer anderen Stelle wieder etwas auf.

Die kurze Laufzeit des Modells ist also nicht die gesamte Arbeitszeit. Wir müssen mindestens unterscheiden zwischen der Entwicklung des Ablaufs, der Erzeugung einer Ausgabe und der anschließenden Prüfung. Ich will hier zunächst nur zeigen: Technisch ist eine solche Ausgabe in wenigen Minuten möglich. Welche Vorbedingungen damit einhergehen und was sie für unsere Forschung wert ist, schauen wir uns noch an.

Ein integriertes Werkzeug am Beispiel von MAXQDA Tailwind

Vielleicht möchten Sie nicht selbst Modelle auswählen und Prompts entwickeln. Dann liegt es nahe, eine Anwendung zu verwenden, die diese Vorarbeiten übernimmt. MAXQDA ist an vielen Hochschulen verfügbar, sei es über Einzel- oder Campuslizenzen. Neben den KI-Funktionen innerhalb der Programmoberfläche gibt es mit Tailwind eine webbasierte Anwendung. (MAXQDA Tailwind)

Ich nehme wieder dieselben Interviews aus unserer Lehrveranstaltungsevaluation. Für die Demonstration lade ich einige davon in Tailwind. Schon beim Hochladen werden Zusammenfassungen erzeugt. Ich muss dafür keine eigene Anweisung formulieren. Im gezeigten Beispiel dauert das ungefähr zwanzig Sekunden.

In den Zusammenfassungen stehen grüne Zahlen. Das sind Verweise auf das Originalmaterial. Wenn ich auf einen solchen Verweis klicke, springt die Anwendung zum passenden Absatz und hebt ihn hervor. Ich kann die Aussage aus der Zusammenfassung also an der entsprechenden Stelle im Original überprüfen.

Damit habe ich bereits Fallzusammenfassungen in beeindruckender Geschwindigkeit erzeugen lassen.

Themen finden und ausarbeiten lassen

Wenn wir an qualitative Inhaltsanalyse denken, sind die einzelnen Fallzusammenfassungen häufig erst ein Anfang. Uns interessieren Themen, die über mehrere Interviews hinweg vorkommen. Auch dafür gibt es einen Button.

In einem zunächst leeren Themenfenster kann ich „Find Topics" wählen. Das Werkzeug arbeitet dann das Material durch und schlägt Themen mit Bezeichnungen und Beschreibungen vor. In meiner Demonstration geschieht das in weniger als zehn Sekunden. Eine eigene Forschungsfrage habe ich an dieser Stelle noch nicht eingegeben.

Bei den Evaluationsinterviews erscheinen Vorschläge wie Lernstrategien und Arbeitsweisen, Bewertung der Lehrformate, Klausurvorbereitung, soziale Kontexte oder Kritik und Verbesserungsvorschläge. Das sind Themen, die mich für eine Lehrveranstaltungsevaluation interessieren könnten. Die Vorschläge beziehen sich erkennbar auf das Material. Nehmen wir beispielsweise einen Themenbereich zur Organisation des Lernens. Er umfasst die individuelle Vor- und Nachbereitung der Lehrinhalte, die Nutzung von Literatur sowie Entscheidungen für Einzel- oder Gruppenarbeit. Solche Beschreibungen wirken nicht einfach wie beliebige Überschriften, die man jedem Datensatz geben könnte. Es sind Aspekte, um die es in diesen Interviews geht.

Wenn ich nun denke: „Ja, darüber möchte ich im Ergebnisbericht etwas sagen“, wähle ich den Vorschlag aus und lasse das Thema erstellen. Dann wird gleichzeitig eine themenbezogene Ausarbeitung erzeugt. Auch das dauert im gezeigten Beispiel nur ungefähr zwanzig Sekunden. Nehmen wir als Beispiel das Thema „Lernverhalten und Strategien“. Oben steht eine Beschreibung des Themas. Darunter folgt die Ausarbeitung, gegliedert beispielsweise nach der Entwicklung des Lernverhaltens im Semesterverlauf und der Nutzung von Lernmaterialien. Dabei werden ganz konkrete Begriffe aus den Interviews aufgenommen. In einem Beispiel wird „der Bortz“ genannt, also ein Statistiklehrbuch von Jürgen Bortz, das damals ein Standardwerk in der Veranstaltung war. Umgangssprachlich sagten Studierende eben „der Bortz“. Die Ausgabe greift diese Bezeichnung auf und bezieht sie auf einen bestimmten Fall.

Auch hier finde ich wieder die Verweise, über die ich zum Originalabsatz springen kann. Ich kann nachsehen, ob die Aussage stimmt und ob die zugeordnete Stelle sie trägt. Wenn mir die Themenbeschreibung nicht passt, kann ich sie verändern oder ein Thema selbst beschreiben. Anschließend wird die Ausarbeitung auf dieser Grundlage erzeugt.

Damit liegt innerhalb kurzer Zeit etwas vor, das wie ein ausgearbeiteter deskriptiver Ergebnisteil aussieht.

Eine Typenbildung anfordern

Zwei weitere Möglichkeiten will ich zeigen. Qualitative Inhaltsanalyse ist nicht auf das deskriptive Zusammenstellen von Themen beschränkt. Bei Kuckartz und Rädiker werden verschiedene Grundformen beschrieben, darunter die typenbildende qualitative Inhaltsanalyse. Wenn das zur Forschungsfrage und zum Material passt, kann eine Typenbildung interessant sein.

Was mache ich dafür in dieser Oberfläche? Ich gehe zum Chat. Der ist nicht ausschließlich für Typenbildung vorgesehen, sondern kann auf sehr unterschiedliche Anfragen reagieren. Man kann eine Metaphernanalyse anfordern, nach Diskursen fragen oder eben eine Typenbildung wünschen.

Ich habe das bewusst hanebüchen vorbereitet. Meine Anfrage lautet sinngemäß: „Mach mir mal eine qualitative Typenbildung und erkläre zu jedem Typus, wie ich auf diesen besser eingehen könnte.“ Sie ist ausgesprochen knapp und methodisch kaum bestimmt. Nach etwa 45 Sekunden kommen nicht nur Beschreibungen von Typen, sondern auch Listen charakteristischer Merkmale und zugeordnete Zitate. Und es kommen genau die Handlungshinweise, um die ich gebeten habe, wie bspw. frühzeitig die Anwendbarkeit betonen, klare Strukturen und Lernpfade anbieten, überschaubare Einheiten bilden und so weiter.

Von einer unbestimmten Frage zum vollständigen Report

Vielleicht sagt jemand, das waren jetzt immer noch mehrere Klicks. Ich möchte einfach einen vollständigen Ergebnisbericht. Auch dafür gibt es eine Funktion.

Für einen neuen Report wähle ich die Texte aus, mit denen ich arbeiten möchte, und gebe eine Forschungsfrage ein. Nun ist es manchmal gar nicht so einfach, eine gute Forschungsfrage zu formulieren. Also tue ich in der Demonstration so, als wäre mir noch

nicht klar, worauf ich hinauswill. Meine Eingabe lautet ungefähr: „Na ja, was man da so rauskriegen kann und dass es wissenschaftlich ist und passt."

Daraufhin kann ich mir alternative Fragen vorschlagen lassen. Das Werkzeug formuliert verschiedene Richtungen, etwa Lernstrategien, Einstellungsentwicklung oder Kompensationsmechanismen. Ich selbst habe diese Begriffe nicht vorgegeben; sie werden mit Bezug auf das Material erzeugt. Ein Vorschlag lautet: „Wie kompensieren Studierende wahrgenommene Defizite in der didaktischen Vermittlung der Vorlesung durch alternative Lernressourcen und soziale Unterstützungssysteme?" Das klingt doch erst einmal ganz gut. Es klingt nach einer wissenschaftlichen Forschungsfrage. Und wenn ich sie auswähle, lässt sich dazu ein Report erstellen.

In der Demonstration dauert das ungefähr dreißig Sekunden. Dann liegt ein Bericht mit einem Abstract vor, der sich an der ausgewählten Frage abarbeitet und sich auf das Material bezieht. Diesen Bericht kann ich exportieren und anschließend weiterbearbeiten.

An dieser Stelle könnte man eigentlich mit dem Vortrag aufhören. Es ist der Horror und die utopischste Glücksfantasie gleichzeitig, je nachdem, wo man sich selbst positioniert. Das Material ist da, die Anwendung erzeugt eine Frage, bearbeitet Themen und liefert einen Bericht. So sieht es jedenfalls zunächst aus.

Fünf Studien zur Nutzung von KI

Bevor wir die methodischen Herausforderungen genauer anschauen, möchte ich fünf Studien einbeziehen. Sie betreffen nicht alle qualitative Forschung. Sie geben uns aber Hinweise darauf, wie Menschen mit KI-Ausgaben umgehen, welche kognitiven Folgen bestimmte Nutzungsweisen haben können und welche Rolle eine strukturierte Gestaltung spielt.

Die erste und die letzte Studie haben für mich die stärkste Symbolwirkung. Die erste zeigt ein Problem der ungeprüften Übernahme. Die letzte weist in eine Richtung, in der wir über Gestaltung nachdenken können.

Falsche Antworten und größere Selbstsicherheit

Shaw und Nave haben in drei Experimenten untersucht, wie Menschen KI in Aufgaben des schlussfolgernden Denkens einbeziehen. Insgesamt nahmen 1.372 Menschen teil; es gab 9.593 Aufgabendurchgänge. Die Teilnehmenden bearbeiteten Knobelaufgaben, bei denen die auf den ersten Blick naheliegende Antwort häufig falsch ist. Sie konnten entscheiden, ob sie ein KI-System konsultieren wollten. (Shaw / Nave 2026)

Die Forschenden haben beeinflusst, ob die KI richtige oder falsche Antworten gab. Besonders interessant sind die Durchgänge, in denen sich Menschen an die KI wandten und diese gezielt falsch antwortete. In ungefähr drei von vier solcher Durchgänge übernahmen die Menschen dennoch die falsche Antwort. Gleichzeitig wuchs die Sicherheit, eine richtige Antwort gegeben zu haben.

Das finde ich bemerkenswert: Nicht nur die falsche Antwort wird übernommen. Hinterher ist die Person auch noch sicherer, richtig zu liegen.

Die Autoren sprechen von „cognitive surrender". Gemeint ist die Übernahme einer KI-Ausgabe mit geringer eigener Prüfung. Die Kompetenz des Systems wird so hoch eingeschätzt, dass die eigene Überlegung zurücktritt. Sinngemäß: Die KI wird es besser wissen als ich.

Weil die Richtigkeit der KI-Antworten experimentell variiert wurde, untersuchen die Autoren innerhalb dieses Versuchs einen ursächlichen Einfluss. Das ist mehr als die bloße Beobachtung, dass Menschen mit hoher KI-Nutzung sich anders verhalten. Es bedeutet aber noch nicht, dass wir denselben Effekt in derselben Stärke für jede wissenschaftliche Aufgabe annehmen können.

Für unseren Zusammenhang ist die Frage wichtig: Wie widerständig verhalten wir uns gegenüber einer sprachlich überzeugenden Ausgabe? Prüfen wir sie noch einmal selbst, oder ist die Tatsache, dass sie vom Modell kommt, bereits ein Grund, sie für richtig zu halten?

Selbstvertrauen und kritische Auseinandersetzung

Die zweite Studie stammt von Lee und anderen. Sie befragten Wissensarbeitende nach ihrer Nutzung generativer KI und nach der dabei erlebten kritischen Auseinandersetzung. Höheres aufgabenbezogenes Selbstvertrauen hing mit mehr selbstberichteter kritischer Auseinandersetzung zusammen. Höheres Vertrauen in die KI hing dagegen mit weniger solcher Auseinandersetzung zusammen. (Lee u. a. 2025)

Das ist nicht dasselbe wie ein objektiver Nachweis, dass Fachleute immer besser prüfen und Anfängerinnen oder Anfänger immer weniger prüfen. Untersucht werden hier unter anderem Selbsteinschätzungen. Die Richtung ist für unsere Überlegung trotzdem interessant: Traue ich mir selbst zu, eine Aufgabe zu bearbeiten und die Antwort zu beurteilen, oder vertraue ich vor allem dem System?

Mir erscheint nachvollziehbar, dass eine eigene begründete Position hilft, widersprechen zu können. Wenn mir die Grundlagen fehlen, muss ich sie mir erst erarbeiten. Sonst kann ich schwer sagen: Moment, das stimmt hier nicht.

Das Gemeine an KI-Ausgaben ist ja, dass wir häufig eine Mischung bekommen. Vieles ist korrekt, plausibel und brauchbar. Einzelne Teile sind aber nicht passend, falsch oder zu weitgehend. Genau diese Stellen zu erkennen, braucht mehr als ein allgemeines Misstrauen gegenüber KI. Es braucht die Fähigkeit, am konkreten Inhalt zu urteilen.

Essays schreiben und sich an den eigenen Text erinnern

Die dritte Studie von Kosmyna und anderen befasst sich mit dem Schreiben von Essays. Die Teilnehmenden arbeiteten entweder mit einem Sprachmodell, mit einer Suchmaschine oder ohne diese Hilfsmittel. Untersucht wurden unter anderem die Texte, die Erinnerung daran und die während der Aufgabe gemessene Hirnaktivität. (Kosmyna u. a. 2025)

Teilnehmende der LLM-Gruppe berichteten ein geringeres Gefühl eigener Urheberschaft. Sie konnten Passagen ihres Textes unmittelbar danach schlechter korrekt zitieren und erinnern. Während der Schreibaufgabe zeigte diese Gruppe außerdem eine schwächere Vernetzung der im EEG erfassten Hirnaktivität.

Die Studie gibt für diese Schreibaufgabe also Hinweise auf eine andere kognitive Beteiligung. Die erzeugten Texte waren den Schreibenden weniger präsent. Das sind Befunde in einer bestimmten Aufgabe, keine allgemeine Aussage über jede Form von KI-Nutzung.

Die Überlegung dahinter finde ich wichtig: Wenn ich eine anstrengende Tätigkeit an ein Werkzeug abgebe, kann ich schneller zu einem Ergebnis kommen. Gleichzeitig ist damit noch nicht dieselbe eigene Auseinandersetzung verbunden. Das wird häufig als „cognitive offloading" beschrieben. Ein Text kann fertig sein, ohne dass ich mich mit seinen Inhalten in derselben Weise beschäftigt habe wie beim eigenen Erarbeiten.

Für die Lehre stellt sich damit eine ziemlich naheliegende Frage: Was soll durch die Bearbeitung einer Aufgabe gelernt werden? Wenn die Aufgabe vor allem dazu dient, selbst etwas zu verstehen und zu entwickeln, dann ist die schnelle Abgabe eines gut aussehenden Ergebnisses nicht der einzige Maßstab.

Möglicher Kompetenzverlust nach der Einführung von KI

Die vierte Studie kommt aus einem anderen Feld, nämlich der Medizin. Budzyń und andere untersuchten die Erkennung von Adenomen bei Darmspiegelungen. Sie verglichen Untersuchungen ohne KI-Unterstützung aus den drei Monaten vor und den drei Monaten nach der Einführung eines KI-Systems in mehreren Zentren. Es handelt sich um eine retrospektive Beobachtungsstudie. (Budzyń 2025)

Nach der Einführung der KI lag die Erkennungsrate bei den Untersuchungen ohne KI niedriger als zuvor. Sie sank von 28,4 auf 22,4 Prozent. Das sind sechs Prozentpunkte beziehungsweise relativ ungefähr 21 Prozent. Die Autoren interpretieren den Befund vorsichtig als einen möglichen Hinweis auf „Deskilling", also eine Abnahme von Fähigkeiten oder verändertes Verhalten nach der Gewöhnung an Unterstützung.

Ich finde diesen Befund als Warnsignal interessant. Gleichzeitig können wir ihn nicht einfach verallgemeinern. Es ist ein Beobachtungsbefund; die Erklärung eines möglichen Kompetenzverlusts bleibt eine vorsichtige Interpretation.

Mich erinnert die mögliche Erklärung an den Sport. Wenn ich über längere Zeit bestimmte Fähigkeiten nicht mehr trainiere, dann kann ich nicht selbstverständlich davon ausgehen, dass ich sie unverändert behalte. Vielleicht gibt es auch bei bestimmten kognitiven Tätigkeiten vergleichbare Fragen. Das ist hier eine Analogie, kein Beweis für den Mechanismus.

Trotzdem lohnt es sich, den Gedanken auf unsere Arbeit zu beziehen: Welche Fähigkeiten benötigen wir weiterhin selbst, und wie erhalten wir sie, wenn uns Systeme einen Teil der Tätigkeit abnehmen? Das kann für erfahrene Personen anders aussehen als für Menschen, die das Verfahren überhaupt erst lernen sollen.

Strukturierte Nutzung als Gestaltungsmöglichkeit

Es gibt aber auch konstruktive Signale. Die fünfte Studie von Gerlich untersuchte mit 150 Teilnehmenden unterschiedliche Formen der Bearbeitung, darunter unangeleitete KI-Nutzung und eine strukturierte Nutzung. Die Argumentqualität wurde anhand von Bewertungen eingeschätzt; hinzu kamen Angaben zur reflektierten Beteiligung. (Gerlich 2025)

Zur strukturierten Nutzung gehören eigene Vorüberlegungen, ein gezielter Einsatz der KI, kritische Prüfung und eine Reflexion über den Prozess. Menschen denken also zunächst selbst und nutzen die KI dann in einem angeleiteten Ablauf. In dieser Bedingung erzielten die Teilnehmenden eine höhere Argumentqualität als bei der unangeleiteten Nutzung.

Das ist für mich ein interessanter Hinweis auf pädagogische Gestaltungsarbeit. Wir sollten allerdings nicht mehr daraus machen, als untersucht wurde. Eine höhere Qualität in diesen Aufgaben ist noch kein allgemeiner Nachweis, dass langfristig mehr gelernt und dauerhaft verinnerlicht wurde.

Aber wir bekommen eine Richtung: Es kommt offenbar darauf an, wie wir die Nutzung gestalten. Der Mensch formuliert erst eine eigene Position, setzt die KI gezielt ein, prüft und reflektiert. Das ist ein anderes Vorgehen als: Hier ist mein Material, gib mir bitte den fertigen Bericht.

Und spätestens damit sind wir bei dem Punkt, den ich im weiteren Vortrag genauer zeigen möchte. Eigene Vorüberlegungen, kritische Prüfung und Reflexion brauchen Zeit. Wenn wir diese Arbeit ernst nehmen, sind wir nicht nach vier Minuten mit Zusammenfassung, Codierung und wissenschaftlicher Auswertung fertig.

Das bedeutet nicht, dass kein Zeitgewinn möglich ist. Es bedeutet, dass wir genauer hinschauen müssen, welche Zeit wir einsparen, welche Arbeit hinzukommt und welche Aufgabe beim Menschen bleibt. Genau diese Differenzierung brauchen wir für die nun folgenden Herausforderungen.

2

TEIL 2

Herausforderungen beim Einsatz von KI

Eine Auswahl der Herausforderungen

A - Menge und Tempo der Entwicklung

B - Das Promptproblem am Beispiel einer Zusammenfassung

C - Das Prüfproblem von LLM-Ausgaben

Im zweiten Block möchte ich die gezeigten Möglichkeiten kritisch einordnen – nicht nur aus meiner eigenen Erfahrung, sondern auch anhand weiterer Studien, Artikel und Positionen. Ich beginne mit einem Überblick über zentrale Herausforderungen.

Eine Auswahl der Herausforderungen

Ich fange mit den bekannten Begriffen an: Bias, Halluzination und Varianz. Dahinter stecken unterschiedliche Probleme, die Auswirkungen auf qualitative Forschung haben. Einige werde ich gleich an kleinen Beispielen zeigen.

Zum Bias habe ich vor Kurzem einen Blogbeitrag von Hughes und Tarrant gelesen. Das beschriebene Testszenario war einfach. Bei zwei Lebenslaufpaaren blieben die Inhalte gleich, verändert wurden Name und Geschlecht der Person. Mit diesen Varianten wurde dasselbe Sprachmodell um Empfehlungen für ein Jahresgehalt und um Argumente für die Gehaltsverhandlung gebeten. Die empfohlenen Beträge und Begründungen unterschieden sich. Auch die erzeugten biografischen Profile variierten deutlich. Ein Gender-Bias ist offensichtlich in diesem kleinen illustrativen Beispiel. (Hughes / Tarrant 2026)

Das fand ich erschreckend. Einen Gender-Bias können wir vermutlich vorab noch recht gut antizipieren und gezielt danach suchen. Es wird aber auch Aspekte geben, an denen uns eine Verzerrung nicht auffällt. Vielleicht weil sie weniger offensichtlich ist, noch nicht untersucht wurde oder wir selbst die entsprechende Perspektive nicht mitbringen. Die Ergebnisse eines Tools können solche Verzerrungen tragen. Dazu kommen die Möglichkeit erfundener Angaben und die Varianz zwischen mehreren Ausgaben.

Eine weitere Herausforderung ist die Tool- und Modellwahl. Welches Angebot benutze ich? Arbeite ich mit Claude, mit GPT oder mit einem Open-Weight-Modell? Wenn ich das selbst

entscheiden möchte, brauche ich eigentlich Testszenarien. Für welche Aufgabe ist welches Modell brauchbar? Wo liegen seine Schwächen? Das herauszufinden kostet Zeit.

Wenn ich selbst Prompts entwickle, kommt eine methodische Entscheidung hinzu. „Mach mir mal eine qualitative Inhaltsanalyse" ist dafür sehr unbestimmt. Wir haben ja nicht die eine Inhaltsanalyse. Es gibt unterschiedliche Vorschläge, etwa von Mayring, Kuckartz und Rädiker, Meuser & Nagel oder Schreier. Und innerhalb dieser Verfahren gibt es verschiedene Facetten und Arbeitsschritte. Unsere Prompts müssen abbilden, was wir davon erwarten und wie wir vorgehen wollen.

Ein Prompt ist deshalb nicht nur etwas Technisches. Er enthält Entscheidungen für bestimmte methodische Arbeitsschritte und für deren Ausgestaltung. Teilweise müssen wir dabei explizieren, was in unserer bisherigen Auswertungspraxis selbstverständlich erschien, aber noch nicht in dieser Form aufgeschrieben war. Prompts transportieren methodische Entscheidungen, ob wir das ausdrücklich beabsichtigen oder nicht.

Damit verbunden ist die noch offene Methodenintegration. Es gibt Vorschläge, und einige stelle ich im dritten Teil vor. Aber daraus folgt noch nicht, dass die Funktionen eines bestimmten Tools bereits selbstverständlich in ein anerkanntes Verfahren eingeordnet wären. Wenn ich in einem QDA-Programm auf einen KI-Button klicke, ist damit noch nicht begründet, welche methodische Aufgabe die Ausgabe in meinem Projekt erfüllt. Neben der Forschungsfrage ist die gewählte Methodik ein Dreh- und Angelpunkt des qualitativen Forschungsprojekts. Sie prägt, welche Daten wir nutzen, welche Transkriptionsregeln angemessen sind, welche Arbeitsschritte aufeinander folgen und wie wir Ergebnisse darstellen. Wer KI einsetzen möchte, muss deshalb im Methodenteil auch begründen können, wie diese Nutzung mit dem eigenen Vorgehen zusammenhängt.

Dann haben wir den Aufwand kritischer Prüfung. Der wird nach meiner Erfahrung erheblich unterschätzt. In meinen eigenen Artikel zu diesem Thema sind fünf Monate Arbeit eingeflossen. Ich hatte gehofft, dass es schneller geht. Gerade darüber wurde mir deutlich, wie groß die Herausforderung ist, wenn ich Ausgaben nicht nur erzeugen, sondern verantwortlich beurteilen möchte.

Nun könnte man sagen, ich entwickle die Prompts gar nicht selbst, sondern verwende die vorhandenen Tools. Dann habe ich allerdings häufig das umgekehrte Problem: Ich sehe die Prompts nicht. Ich sehe nicht vollständig, welche Instruktionen hinter einer Funktion stehen und welches Modell sie auf welche Weise ansprechen. Damit kann ich auch die darin enthaltenen methodischen Entscheidungen nur eingeschränkt reflektieren. Passt die Perspektive zu meiner Forschungsfrage? Werden bestimmte Aspekte bevorzugt? Was wird gewichtet, was ausgelassen? Und wie dokumentiere ich dieses Vorgehen, wenn weder die vollständige Instruktion noch die konkrete Modellversion nachvollziehbar vorliegt? Ein Tool wie MAXQDA oder Qinsights ist ja selbst kein Sprachmodell. Es ist unter anderem eine Benutzeroberfläche mit bestimmten Funktionen und Instruktionen, die ein Modell ansprechen. Für die wissenschaftliche Dokumentation reicht deshalb die Angabe „Ich habe die KI MAXQDA genutzt" nicht aus, um den tatsächlichen Prozess zu erklären. Interessant sind die konkrete Funktion, die Instruktion, das verwendete Modell und das, was ich selbst mit der Ausgabe gemacht habe.

Eine weitere Frage betrifft die Lehre. Wenn Menschen, die noch nie qualitativ geforscht haben, solche Aufgaben automatisieren lassen, wie lernen sie dann, was qualitative

Forschung bedeutet? Wie entwickeln sie die Fähigkeiten, die Ausgaben zu beurteilen? Und was vermittle ich im Rahmen der Methodenausbildung?

Für Studierende und Promovierende kommt die Frage nach der Eigenleistung hinzu. Bisher war die Auseinandersetzung mit dem Datenmaterial, die Entwicklung von Erkenntnissen und deren Darstellung ein zentraler Teil der forschungsprägenden Eigenleistung. Wenn eine KI davon einen großen Anteil übernimmt, muss deutlich werden, worin die eigene Leistung besteht. Und wo übernehme ich sichtbar die Agency, also die Steuerung und Verantwortung? Wie zeige ich, dass es weiterhin eine von mir geleitete Forschung unter Verwendung von KI ist? Und ist das in der betreffenden Qualifikationsarbeit überhaupt erlaubt? Die geltenden Regelungen und der konkrete Umgang mit KI müssen geklärt werden. Ich sollte das nicht erst dann versuchen, wenn ich die Daten schon verarbeitet habe und den Ergebnisbericht abgeben möchte. Ein konkretes Beispiel für mögliche Folgen gibt es aus Kassel. Das Verwaltungsgericht hat am 25. Februar 2026 in zwei Fällen die Bewertung von Prüfungsarbeiten als nicht bestanden und den Ausschluss von der Prüfungswiederholung wegen besonders schwerer Täuschung bestätigt. Es ging um nicht offengelegte KI-Hilfe (Verwaltungsgericht Kassel 2026).

Damit sind die Herausforderungen noch lange nicht erschöpft. Es gibt ökonomische Fragen: Wer hat Zugang und wie wird das finanziert? Ethische und ökologische Fragen: Welche Folgen hat die Nutzung, welchen Energie- und Ressourcenverbrauch haben die Rechenzentren? Rechtliche und politische Fragen: Welche Abhängigkeiten und Machtverhältnisse entstehen? Ich werde diese Themen hier nicht alle ausführen, aber sie gehören dazu.

Eine weitere Herausforderung ist das Tempo der Entwicklung. Ich beschäftige mich seit rund zwei Jahren jeden Tag mit Sprachmodellen, probiere etwas aus und lese dazu. Ich habe immer wieder Phasen, in denen ich denke, jetzt wird mir das zu viel. Wie ist das erst für Menschen mit anderen Arbeitszusammenhängen, die gar nicht täglich die Zeit dafür haben? Da entsteht schnell das Gefühl, abgehängt zu werden oder viel nacharbeiten zu müssen.

All das trägt zu einer uneinheitlichen Akzeptanz bei. Neben neugierigen Anwendungsexperimenten gibt es deutliche Ablehnung. Jowsey und andere haben sich beispielsweise gegen den Einsatz generativer KI in reflexiver qualitativer Forschung positioniert. (Jowsey u. a. 2025) Wir haben es mit einer Gemengelage aus Experimenten, Abwehr und laufender Entwicklung zu tun. Dass Anwendungen vorhanden sind und genutzt werden, bedeutet noch nicht, dass ihre wissenschaftliche Anerkennung geklärt wäre.

Gleichzeitig kommen Menschen an die Hochschulen, für die KI-Anwendungen bereits ein alltägliches Arbeitswerkzeug sind. Das erzeugt einen erheblichen Spagat. Die Nutzung ist schon da, während wir als Forschende und Lehrende noch daran arbeiten, sie methodisch zu durchdenken und institutionell zu gestalten.

Aus dieser Auswahl an Herausforderungen greife ich nun drei Bereiche heraus: erstens Menge und Tempo der Entwicklung, zweitens die Promptgestaltung und drittens die kritische Prüfung.

A - Menge und Tempo der Entwicklung

Bei der Menge reicht ein Blick auf vier neue Bücher allein im Oktober 2026. Zusammen kommen sie auf rund tausend Seiten: *Qualitative Data Analysis with AI*, *KI in der qualitativen Forschung*, *Interpretieren zusammen mit KI* und *Qualitativ Forschen mit KI über KI*. Die Titel klingen teilweise ähnlich, die Zugänge unterscheiden sich aber.

Der von Susanne Friese und David Morgan herausgegebene Sammelband *Qualitative Data Analysis with AI* versammelt unterschiedliche Vorschläge, wie KI in qualitativer Forschung eingesetzt werden könnte. Friese ist im deutschsprachigen Raum mit dieser Thematik und ihrem Tool Qinsights sehr präsent; Morgan ist seit Langem in der qualitativen Forschung tätig. In dem Sammelband haben wir auch einen Beitrag zum hybriden Interpretieren veröffentlicht. (Friese / Morgan 2026)

Kempny und andere führen den Einsatz von KI entlang des Forschungsprojekts von der Studienplanung bis zur Datenauswertung aus. Beim ersten Durchsehen ist mir eine ausgeprägte kritische Perspektive nicht sofort aufgefallen. (Kempny u. a. 2026)

Qualitativ Forschen mit KI über KI, herausgegeben von Hella von Unger, Yves Jeanrenaud und Thomas Raucheger, ist aus einem soziologischen Lehrforschungsprojekt an der LMU hervorgegangen. Darin geht es unter anderem um KI-geführte Interviews mit Studierenden und die Reflexion dieser Erfahrungen. Als Experimentierfeld finde ich das sehr spannend. (von Unger / Jeanrenaud / Raucheger 2026)

Schließlich das Buch von Burkhard Schäffer, Fabio Roman Lieder und Franz Krämer: *Interpretieren zusammen mit KI*. Schäffer kommt aus der dokumentarischen Methode, also einem rekonstruktiven Ansatz. Die Autorengruppe hat früh zur Arbeit mit Sprachmodellen und zur Promptentwicklung publiziert. Aus diesem Buch greife ich gleich einige Überlegungen auf. (Schäffer / Lieder / Krämer 2026)

Das ist eine Menge. Wer diese Bücher gründlich durcharbeiten und die Vorschläge kritisch diskutieren oder die eigene Arbeitspraxis daran ausrichten möchte, braucht Zeit. Und währenddessen kommen schon wieder neue technische Möglichkeiten hinzu.

Für das Tempo möchte ich ein Beispiel geben, das mich selbst überrascht hat. Es geht um die Möglichkeit, mit KI eine Software nachzubauen. Das geht seit ein paar Tagen mit Google Gemini 4 Argon. Beispiele dazu (Photoshop Suite) kursieren inzwischen zahlreich. Ich wollte das anhand eines Beispiels ausprobieren, das ich selbst gut beurteilen kann. Deshalb habe ich die Software genommen, an deren Entwicklung wir seit vielen Jahren arbeiten: f4 2026. Ich habe Gemini 4 den Link zum Handbuch gegeben und einen Screenshot. Nach ungefähr 15 Minuten hatte ich eine brauchbare Oberfläche, die f4 zu fast 100 % ähnlich sah und in der die gezeigten Funktionen tatsächlich bedienbar waren. Ich konnte Fenster ein- und ausblenden, Text markieren, Codes unterschiedliche Farben geben, Memos schreiben und Codes verschieben, Retrieval und Suchfunktionen durchführen, Audiodateien abspielen u. v. m. Besonders überraschend war für mich, dass diese Oberfläche im Browser lief: Ich konnte die Funktionsnachbildung direkt auf einem iPad nutzen – eine Umsetzung, die wir zuvor aus Kostengründen nicht einfach nebenbei realisiert hätten. In dieser Vorführung war sie nach einer Viertelstunde benutzbar. Das reichte, um mich zu fragen, wie sich die Entwicklung und Bereitstellung von Software verändern wird. Vielleicht können Menschen künftig viel häufiger sagen: Ich brauche diese und jene Funktion, erstelle mir eine dafür passende

Oberfläche. Kombinieren Sie das mit dem, was wir zuvor in Tailwind gesehen haben. Dort werden bestimmte Auswertungsaufgaben in kurzer Zeit automatisiert. Hier wird eine Oberfläche erzeugt, mit der ich selbst arbeiten kann. Das wirft auch Fragen danach auf, welche Leistungen wir künftig einkaufen und welche wir uns für einen bestimmten Zweck selbst erzeugen lassen.

Eine weitere Entwicklung betrifft die Erzeugung neuer wissenschaftlicher Ergebnisse. OpenAI hat im Herbst 2026 mathematische Manuskripte und Beweisartefakte eines internen Modells veröffentlicht. Solche Erfolgsmeldungen werfen die Frage auf, wie sich die Aufgaben von Forschenden verändern, wenn KI zu bisher offenen Problemen Beiträge liefert. Die Szene der Mathematikwissenschaften ist davon emotional stark betroffen und sieht sich einigen Herausforderungen gegenüber (OpenAI 2026).

Ich möchte daraus keinen direkten Beweis ableiten, dass eine KI deshalb auch die bessere qualitative Forschungsfrage stellt und die automatische Analyse und Schlussfolgerung gleich mit. Aber die Frage bleibt: Welche Aufgaben übernehmen wir als Forschende, wenn Systeme nicht nur vorhandene Texte zusammenfassen, sondern weitergehende Vorschläge und Ergebnisse erzeugen? Mit solchen Entwicklungen müssen wir uns auseinandersetzen.

B - Das Promptproblem am Beispiel einer Zusammenfassung

Kommen wir zur zweiten Vertiefung. Wir nehmen einen der scheinbar einfachsten Fälle und etwas, das viele bereits mit KI gemacht haben, eine Zusammenfassung. Auf der Folie steht folgende Passage aus dem Statistikzusammenhang:

> *„Am Anfang hatte ich echt Horror vor Statistik, ich hab in der Vorlesung gar nichts verstanden. Dann hab ich mich jedes Wochenende ein paar Stunden drangesetzt. Jetzt am Ende komm ich eigentlich ganz gut mit. Aber so richtig wohl fühl ich mich damit immer noch nicht."*

Der Prompt, der verwendet wurde, lautete:

> *„Erzeuge eine rein deskriptive Zusammenfassung"*

Die auf der Folie dokumentierte Ausgabe von Google Gemini 3.8 Flash vom 24. Oktober 2026 lautet:

> *„Trotz anfänglicher Überforderung führte regelmäßiger Eigenaufwand zu soliden Fortschritten, auch wenn eine grundlegende Unsicherheit bestehen blieb."*

Da würden wir vielleicht zunächst sagen, das ist doch gut zusammengefasst. Die wesentlichen Aussagen scheinen enthalten zu sein. Schauen wir genauer hin, sehen wir aber Verschiebungen. Es werden andere Begriffe verwendet. Aus der zeitlichen Abfolge wird ein ursächlicher Zusammenhang. Der Eigenaufwand „führte" zu Fortschritten. Das „eigentlich ganz gut" wird zu „soliden Fortschritten". Und das Unbehagen erscheint nun als „grundlegende Unsicherheit".

Man kann darüber diskutieren, wie schwer diese Verschiebungen wiegen. Ich möchte daran zunächst deutlich machen, dass die Anweisung „rein deskriptiv" noch nicht festlegt, was ausgewählt, miteinander verknüpft und gewichtet wird. Wenn wir eine für unser Projekt passendere Zusammenfassung möchten, müssen wir uns mit diesen Entscheidungen beschäftigen. Das ist nicht nur meine eigene Erfahrung. Auch Schäffer, Lieder und Krämer arbeiten heraus, wie viele Entscheidungen in der Promptgestaltung stecken (Schäffer / Lieder / Krämer 2026). Ich möchte das am selben Beispiel mit einem anders formulierten Auftrag verdeutlichen.

Der Prompt muss zur eigenen Fragestellung passen. Die Fragestellung bestimmt mit, aus welcher Perspektive eine Zusammenfassung hilfreich ist. Das Modell kann meine Perspektive nicht einfach wissen. Ich muss sie ihm mitteilen. Nehmen wir an, meine Frage wäre: Wie gehen Studierende mit Lernschwierigkeiten um? Dann könnte der Prompt lauten:

> *„Fasse für jeden Fall zusammen, wie die befragte Person mit Lernschwierigkeiten in der Statistikveranstaltung umgeht. Berücksichtige, welche Schwierigkeiten sie beschreibt, was sie dagegen unternimmt, wie sich ihre Lage im Semesterverlauf verändert und wie sie ihre Situation am Ende einschätzt. Gib die Aussagen als Schilderung der Person wieder und übernimm Wertungen, Einschränkungen und Gefühlsäußerungen möglichst in ihren eigenen Worten. Stelle zeitliche Abfolgen dar, ohne ursächliche Zusammenhänge herzustellen, die die Person nicht selbst nennt. Vermerke, wenn sie sich zu einem dieser Aspekte nicht äußert."*

Bitte verstehen Sie das nicht als den perfekten Zusammenfassungsprompt. Es ist eine weitere Annäherung. Der Unterschied ist, dass hier bestimmte Entscheidungen sichtbar getroffen werden. Welche Schwierigkeiten interessieren mich? Welche Handlungen? Welche Veränderungen? Wie gehe ich mit Wertungen um? Und soll das Modell auch markieren, wenn zu einem erwarteten Aspekt keine Aussage vorliegt?

Wenn wir solche Prompts selbst formulieren, geraten wir in eine Reflexionsschleife. Was will ich eigentlich wissen? Das ist manchmal gar nicht so klar. „Ich will doch einfach nur eine Zusammenfassung" beantwortet diese Frage leider noch nicht. Soll sie der Chronologie folgen? Soll sie sich an meinen Forschungsfragen orientieren? Sollen auch Dinge sichtbar werden, die im Vergleich zu anderen Fällen nicht genannt werden? Welches Raster lege ich an? Wird aus der Ich-Perspektive formuliert oder aus einer Außenperspektive? All das verändert, wie wir die Daten wahrnehmen und welche Aspekte uns für die Entwicklung einer Erkenntnis zur Verfügung stehen. Insofern kann es sehr produktiv sein, Prompts selbst zu entwickeln. Wir müssen dabei Erwartungen explizieren, die unsere Auswertungspraxis möglicherweise bisher unausgesprochen geleitet haben.

Ich merke das an meinen eigenen Formulierungen. „Beschreibe die zentrale Geschichte der Interviews" ist eine andere Aufforderung als „Prüfe, ob sich eine gemeinsame Geschichte begründen lässt". Aber auch die vorsichtigere zweite Fassung gibt „Geschichte" schon als Suchrichtung vor. Die Promptentwicklung zwingt mich, solche Vorannahmen wahrzunehmen.

Einen Prompt entwickeln heißt auch testen und überarbeiten

Mit einer in meinen Augen guten Instruktion ist die nächste Herausforderung noch nicht gelöst. Der Prompt hält zunächst meine Erwartungen fest, aber das garantiert nicht, dass das Sprachmodell genau das tut, was ich beabsichtigt habe. So kann es passieren, dass wir einen umfassenden, sinnvoll erscheinenden Auftrag formulieren und eine Ausgabe bekommen, in der Zitate fehlen, oder die angeblich wörtlichen Zitate paraphrasiert sind. Oder es wurden Inhalte vergessen und Aspekte nicht beachtet. Bei umfangreichen Datenmengen kann gerade bei lokaler LLM Nutzung beispielsweise der Anfang oder das Ende stärker berücksichtigt sein als das Material in der Mitte. Das wird unter dem Stichwort „Lost in the Middle" untersucht. (Liu u. a. 2024)

Deshalb müssen wir den Prompt testen. Wenn wir dabei Schwächen finden, müssen wir ihn verändern und erneut testen. Das ist eine iterative Entwicklung. In alten Vorträgen habe ich hier vor allem meine eigene Erfahrung beschrieben. Seit Oktober 2026 kann ich sie mit methodischen Vorschlägen aus der Literatur unterfüttern. Ich habe dazu die Überlegungen von Schäffer, Lieder und Krämer (vgl. S. 60 – 77; 197 – 200) gebündelt. Vier Entwicklungsphasen lassen sich darin unterscheiden. (Schäffer / Lieder / Krämer 2026)

Zuerst entwerfe ich eine erste Promptfassung und erprobe sie explorativ. Ich richte Rolle, Begriffe, Arbeitsschritte, Grenzen, erwartete Ausgabe sowie theoretisch-methodologische Vorgaben am Forschungsdesign aus. Gegebenenfalls brauche ich Beispiele. Dann prüfe ich: Ist die Umsetzung für meine Aufgabe überhaupt methodisch brauchbar? Was fehlt? Welche Bausteine muss ich gezielt überarbeiten? Danach habe ich eine erste Fassung, aber noch keinen für jede Verwendung abgesicherten Prompt. Im Buch werden umfangreiche Prompts gezeigt, die sich über mehrere Seiten erstrecken. Das macht sichtbar, wie viel Wissen und wie viele Entscheidungen in einer solchen Instruktion stecken können.

Im zweiten Bereich lege ich Testmaterial und Bewertungsmaßstäbe fest. Ich brauche unterschiedliche Interviewsequenzen, an denen sich die Brauchbarkeit beurteilen lässt. Die Autorengruppe schlägt zwölf methodologienübergreifende Kriterien vor, etwa Textnähe, Kontextsensitivität und argumentative Plausibilität. Solche Kriterien muss ich für die Bewertung operationalisieren: Was gilt jeweils als erfüllt? Auf welcher Ebene urteile ich? Welche Abstufungen unterscheide ich? Wenn mehrere Menschen bewerten, brauchen sie eine gemeinsame Grundlage und müssen diese Anwendung erproben.

Im dritten Bereich erzeuge ich Testausgaben und beurteile sie am Original. Dabei suche ich nach methodischen Fehlpassungen, missachteten Anweisungen, erfundenen Aussagen und übersehenen Informationen. Um die Übereinstimmung von Bewertungen zu prüfen, können mehrere Personen dieselben Ausgaben beurteilen. Auch ein Evaluationsmodell kann dafür eingesetzt werden, aber dann muss ich wiederum prüfen, wie dessen Urteile zu menschlichen Einschätzungen passen.

Schließlich werte ich die Bewertungen aus, ziehe Konsequenzen und dokumentiere das Vorgehen. Was habe ich geändert, mit welchem Modell und welchen Einstellungen gearbeitet? Ein Anweisungsproblem führt mich zurück zur Promptgestaltung. Ungeeignetes Testmaterial oder unklare Maßstäbe führen mich zurück zur Festlegung des Prüfdesigns. Ich bin also in einer Schleife aus Erproben, Beurteilen und Überarbeiten.

Nicht einfach vorgegeben sind damit verbindliche Erfolgsschwellen, akzeptable Fehlerquoten und der Umfang einer erneuten Prüfung. Auch dafür müssen wir Entscheidungen treffen und sie begründen.

Wie viel Zeit steckt darin? Ich habe mir angesehen, welche Arbeitsschritte und Dokumentationsbeispiele beschrieben werden, und dafür eine eigene Planungsrechnung aufgestellt. Für ein solches methodisches Entwicklungs- und Evaluationsprojekt, das einen übertragbaren und systematisch geprüften Analyseablauf absichert, komme ich auf 250 bis 500 Arbeitsstunden für einen so entwickelten Prompt. Würde man das nicht so umfassend machen, bleibt für mich eine offene Frage, wie verantwortliches Arbeiten mit KI bei größeren Datenmengen ohne einen derart sorgfältig entwickelten und getesteten Prompt begründet werden könnte. Nicht ohne Grund haben Schäffer, Lieder und Krämer diesen Prüfaufwand in ihrem Buch so ausführlich dargestellt.

Da könnte man sagen, das macht doch niemand, das ergibt für die allermeisten doch keinen Sinn, weil schlicht die Zeit und das Know-how fehlt. Einige tun das allerdings sehr intensiv, nämlich Softwarehersteller, die KI-Funktionen in ihre Produkte integrieren. Die Entwicklung und Prüfung solcher Prompt-Instruktionen ist zentraler Bestandteil ihrer Produktintegration. Leider werden diese Entwicklungsprozesse und die Qualitäts- und Prozessentscheidungen in der Regel nicht transparent geteilt und sind daher nicht nachvollziehbar. Da die Prompts zudem in der Regel nicht offen einsehbar sind, entfällt auch die Möglichkeit der nachträglichen Reflexion derselben.

Für uns entsteht daraus ein Dilemma. Verwenden wir die vorgefertigte Funktion, können wir ihre methodische Ausgestaltung möglicherweise nicht vollständig reflektieren. Wollen wir die Instruktion selbst entwickeln und systematisch prüfen, brauchen wir entsprechende Zeitressourcen und Kompetenz, die deutlich über das hinausgehen, was Menschen im Rahmen von Qualifikationsarbeiten zur Verfügung steht.

C - Das Prüfproblem von LLM-Ausgaben

Damit komme ich zur dritten Vertiefung. Muss ich eigentlich prüfen, was die KI ausgibt?

Eine konkrete Position dazu formuliert Stefan Rädiker in seinem Leitfaden zu MAXQDA AI Assist. Nicht jeder Chatverweis müsse angeklickt und nicht jede Zusammenfassung Satz für Satz am Original geprüft werden. Er empfiehlt, zu Beginn einzelne Ausgaben stichprobenartig abzugleichen und so Erfahrungen mit den eigenen Daten zu sammeln. Gleichzeitig verlangt er eigene Materialkenntnis sowie Fach- und Methodenkompetenz. Zitate vor der Veröffentlichung sowie Literatur- und Theorieverweise sollen immer geprüft werden. (Rädiker 2026, S. 44, 46 und 54)

Meine Rückfrage ist: Woran entscheide ich, dass meine Materialkenntnis und der punktuelle Abgleich die konkrete Aussage hinreichend absichern? Um sich zunächst im Material zu orientieren, kann ich mit einer vorläufigen Zusammenfassung arbeiten. Wenn ich eine Aussage als Forschungsergebnis übernehme, muss ich sie angemessen absichern. Gerade bei einer Zusammenfassung über viele Interviews zeigen passende Einzelstellen noch nicht, ob Auswahl und Gewichtung insgesamt stimmen. Wichtige Gegenpositionen können fehlen.

Konkrete Zahlen nennen Sam Belkin und Kollegen im AQUATIC-Protokoll für schnelle deskriptive Analysen in der Gesundheitsforschung. Pro Forschungsfrage oder Prompt sollen

mindestens 30 Prozent der KI-Antworten im Kontext geprüft werden. Scheitern mehr als zehn Prozent der geprüften Antworten an einer Kontrolle, wird auf vollständige Antwortprüfung erweitert. Zusätzlich sind sämtliche Belegzitate zu prüfen; vorgesehen sind außerdem vorherige Materiallektüre und die Suche nach Gegenbeispielen. Es geht also um ein Bündel von Sicherungen. (Belkin u. a. 2026, S. 2 und 4–6)

Das löst in mir aber mehr Fragen aus, als es Antworten gibt. Was begründet gerade diese Schwellen? Wie wird die Stichprobe ausgewählt? Welche Fehler werden erfasst, welche möglicherweise übersehen? Welche Folgen haben sie für das Ergebnis? Eine solche Schwelle wird nicht dadurch wissenschaftlich begründet, dass eine umfassendere Prüfung zeitaufwendig ist. Mein Einwand richtet sich gegen einen unbegründeten Schluss von begrenzter Kontrolle auf hinreichende Absicherung. Das betrifft wissenschaftliche Redlichkeit und die Grundlagen, an denen wir uns etwa mit den Leitlinien der DFG oder den DeGEval-Standards orientieren. (DFG 2019) (DeGEval 2017)

Warum ist diese Prüfung überhaupt so schwierig? Ich illustriere das an drei Beispielen: dem finalen Charakter einer Ausgabe, ihrer Varianz und dem Verlust relevanter Einzelheiten.

Eine mögliche Lesart erscheint als feststehende Deutung

Hier ist eine kurze fiktive Interviewsequenz, die ich als Lehrbeispiel verwende:

> ***Interviewer:*** *„Zum Anfang: Wie lange sind Sie genau schon bei move35 hier in Marburg dabei?"*
>
> ***Befragte:*** *„Das erste Treffen hat im September stattgefunden. Das war '23. Und seit diesem Punkt sind sie auch ganz aktiv."*

Das Wort „sie" ist hier ein möglicher Stein des Anstoßes. Warum antwortet die Person nicht mit „ich" oder „wir"? Die Frage scheint sich zunächst auf die eigene Beteiligung zu beziehen. Eine mögliche Lesart wäre, dass sich in der Antwort eine Distanz zwischen der Person und einer Gruppe zeigt. Wer rekonstruktiv arbeitet, könnte diese Spur weiterverfolgen. Ob sie trägt, muss sich an weiterem Material zeigen. Nun bekommt Gemini dieselbe kleine Passage mit der Aufforderung, sie zusammenzufassen. Die dokumentierte Ausgabe vom 26. Mai 2026 lautet:

> *„Die Befragte erklärt, seit der initialen Zusammenkunft im September 2023 ununterbrochen und mit hohem Einsatz tätig zu sein."*

Ich würde für diese Zusammenfassung nicht ohne Weiteres unterschreiben. Genauer gesagt: Sie legt eine mögliche Lesart fest, ohne die Mehrdeutigkeit sichtbar zu machen. Die Befragte selbst wird nun als die aktive Person beschrieben. Aus „ganz aktiv" wird außerdem „ununterbrochen und mit hohem Einsatz".

Das Schwierige an Textdaten ist, dass wir es nicht einfach mit Eindeutigkeit zu tun haben. Sprache ist kontextabhängig. Wir arbeiten mögliche Lesarten heraus und prüfen, welche wir am plausibelsten am Material begründen können.

In der Zusammenfassung sehen wir diese Offenheit nicht mehr. Die Aussage ist so formuliert, als sei das die feststehende Bedeutung. Ob wir ihr folgen, lässt sich aber nicht nur anhand ihrer glatten Formulierung entscheiden. Dafür müssen wir zurück in die Daten,

und selbst dann bleibt es schwierig. Die Stelle dieses Beispiels überhaupt als interpretationsbedürftig zu bemerken, braucht Aufmerksamkeit und Erfahrung. Dann brauchen wir Zeit und Fähigkeit, uns mit den möglichen Bedeutungen auseinanderzusetzen und uns gegenüber der fertigen Zusammenfassung auch widerständig zu verhalten.

Varianz zwischen mehreren Ausgaben

Bleiben wir bei derselben kleinen Sequenz. Diesmal lautet die Frage: „Identifiziere Themen." Das wäre ein plausibler Auftrag, wenn ich thematisch codieren möchte. Die erste Anfrage liefert die Vorschläge „Zugehörigkeit und Kontinuität" und „Aktivitätsgrad". Da könnte ich sagen, ja, darüber können wir sprechen. „Ganz aktiv" steht tatsächlich in der Passage. Kurz darauf liefert dieselbe Anfrage die Vorschläge „Historischer Kontext" und „Engagement". Die Bezeichnungen liegen teilweise nahe beieinander, aber sie setzen unterschiedliche Akzente. Welche Vorschläge sind für meine Fragestellung hilfreich? Sollte ich sie zusammenbringen? Brauche ich alle vier? Wie oft müsste ich fragen, um die Bandbreite möglicher Themen zu sehen? Und nach welchem Maßstab wähle ich aus?

Diese Varianz wird mir nicht automatisch deutlich, wenn ich in einer Software einmal auf einen Button drücke. Dann sehe ich eine oder zwei Ausgaben. Ich muss bereits wissen, dass weitere Ausgaben anders ausfallen können und dass ich ihre Bedeutung beurteilen muss. Mehrere Durchläufe können mir helfen, solche Schwankungen wahrzunehmen. Eine feste Zahl an Wiederholungen garantiert aber keine vollständige Erfassung aller relevanten Perspektiven. Und danach bleibt die Frage, welche davon ich aus welchen Gründen übernehme.

Bei umfangreicherem Material kann das erheblichen Aufwand bedeuten. Zugleich sind die Vorschläge häufig so plausibel formuliert wie die Themen, die wir vorhin in Tailwind gesehen haben. Verhalte ich mich dann noch widerständig? Oder übernehme ich die erste brauchbar klingende Ordnung?

Korrekte Einzelbelege sichern keine korrekte Gesamtaussage

Das dritte Problem finde ich besonders wichtig. Eine Ausgabe kann korrekte Zitate und Bezüge enthalten und trotzdem eine entscheidende Gegenposition übergehen. Dann kann die Gesamtaussage unangemessen sein, obwohl die angeführten Einzelbelege stimmen. Nehmen wir noch einmal die Lehrevaluation. Dort sagt die Person B4:

> *„Also das Tutorium, das fand ich langweilig, weil ich da auch nur einmal war, das war wie gesagt nichts für mich, […] aber ich habs halt lieber in solchen Arbeitsgruppen, wo man halt sozusagen gleichberechtigt ist, ohne da jetzt den Tutorien irgendwie was schlechtes anzureden." (B4,* Abs. 12)

Die Person findet das Tutorium für sich nicht passend und zieht eine Arbeitsgruppe vor. Zugleich relativiert sie ihre Kritik. Wenn ich nun das Modell befrage, wie die Tutorien bewertet werden, bekomme ich in der dokumentierten Ausgabe eines lokalen Modells (GLM 4.5 Air):

> *„Tutorien wurden dagegen durchgängig positiv bewertet. B1 betonte: ‚Das Tutorium war noch besser als die Übung, weil da wurden ja wirklich die einzelnen Sachen noch mal durchgegangen', während B9 hervorhob: ‚Was für mich mit am Wichtigsten ist, ist eigentlich das Tutorium.'"*

Danach endet die Darstellung. Die beiden angeführten Quellen sind korrekt. Aber B4 fehlt, einschließlich der dort formulierten Ambivalenz. Gerade die Behauptung „durchgängig positiv" ist deshalb nicht angemessen. Wäre B4 einbezogen worden, hätte ich vielleicht gesagt: Für viele ist das Tutorium hilfreich, aber nicht alle profitieren davon. Was brauchen die anderen? Vielleicht wäre eine begleitete Phase zur Gruppenbildung sinnvoll. Vielleicht müsste man die Gestaltung des Tutoriums verändern. Die ausgelassene Position könnte also zu anderen praktischen Konsequenzen führen.

Eine Zusammenfassung oder ein kurzer Ergebnisbericht trifft zwangsläufig eine Auswahl. Die Ausgabe wird nicht so lang sein wie das gesamte Ausgangsmaterial. Aber ich sehe der Auswahl nicht an, was dabei verloren gegangen ist und warum. Beim Lesen der beiden korrekten Zitate fällt die fehlende Gegenstelle nicht auf. Ich würde das Problem erst erkennen, wenn ich mich selbst mit dem Material in seiner Breite auseinandergesetzt habe: Was steht in den einzelnen Fällen? Wo sind Unterschiede? Welche Gegenpositionen gibt es? Was gehört zusammen? Auf dieser Grundlage kann ich dann sagen: Hier greift die Darstellung zu kurz, hier fehlt etwas Entscheidendes.

Das ist mehr als das Nachschlagen von Zitaten. Es reicht nicht, dass ich prüfe, ob B1 und B9 den jeweiligen Satz wirklich gesagt haben. Ich muss auch beurteilen können, ob ihre Aussagen die gezogene Gesamtschlussfolgerung tragen.

Deshalb unterscheide ich zwei Prüfaufgaben. Erstens den Belegabgleich: Stimmen Wortlaut, Fundstelle und Fallzuordnung? Dafür kann ich eine ausgegebene Angabe mit der Originalquelle vergleichen, auch wenn ich das Material zuvor noch nicht vollständig gelesen habe.

Zweitens die Prüfung auf Gegenstellen, Glättungen, Auslassungen und inhaltlich angemessene Deutung: Welche Erfahrungen fehlen? Welche Unterschiede sind eingeebnet? Lässt sich eine andere Sichtweise materialnah entwickeln? Was wurde unangemessen interpretiert? Dafür brauche ich eine eigene Auseinandersetzung mit den Daten. Ich muss Material erschließen, Fälle vergleichen und Zusammenhänge sowie Unterschiede selbst herausarbeiten.

Mit wachsender Datenmenge steigt auch dieser Aufwand. Wenn ich Ergebnisse ohne eine solche Prüfung übernehme, trage ich Verantwortung für eine Auswahl, Gewichtung und Deutung, deren Angemessenheit ich nicht beurteilen kann. Die Entscheidung ist dann faktisch vom Modell vorgegeben, auch wenn ich sie anschließend als meinen Ergebnistext verwende.

Was lässt sich durch weitere Instruktionen verbessern

Ich könnte das Modell beispielsweise auffordern: Liste zusätzlich auf, welche Aspekte du nicht einbezogen hast, und erläutere, warum. Auch diese Erläuterungen müssen gelesen und bewertet werden. Aber auch dann ist nicht automatisch sichtbar, ob die Liste vollständig ist. Je umfangreicher das Material, desto umfangreicher wäre eine solche Prüfung und Dokumentation. Zu einem kurzen Ergebnisbericht kämen unter Umständen viele Seiten mit Belegstellen und Begründungen hinzu.

In traditioneller qualitativer Auswertungspraxis arbeiten sich Forschende selbst durch das Material, entwickeln ein Verständnis seiner Bandbreite und treffen darauf bezogen eine begründete Auswahl für die Darstellung. Es wird versucht, den Forschungsprozess intersubjektiv nachvollziehbar zu machen. Genau diese Nachvollziehbarkeit wird durch eine fertige LLM-Ausgabe nicht von allein hergestellt. Das Ergebnis liegt zwar als Text vor, doch bleibt darin unsichtbar, was ausgelassen wurde, weshalb eine bestimmte Deutung Vorrang erhielt und welche Konsequenzen eine andere Auswahl gehabt hätte. Das beeinflusst potenziell unsere Ergebnisse.

Die Frage lautet deshalb: Wann bin ich in der Lage, eine bestimmte Arbeitsaufgabe verantwortlich an die KI zu übertragen, und wie muss ich das Ergebnis prüfen? Dafür müssen wir eine begründete Position finden. Ich kann hier keine für jeden Fall fertige Antwort vorgeben. Aber es ist weniger trivial, als wir es uns vielleicht wünschen.

Sie merken, ich bewege mich zwischen den Positionen. Ich bin technisch neugierig, probiere die Werkzeuge aus und lote ihre Möglichkeiten aus. Zugleich sehe ich diese Herausforderungen und möchte sie benennen. Für einige sehe ich noch keine befriedigende Lösung, auch wenn gleich einige methodische Vorschläge folgen.

Was ich aber beobachte: Das Feld nutzt die Funktionen bereits. Ein Kollege aus den USA, der ein qualitatives Forschungszentrum leitet, schilderte mir in einem persönlichen Gespräch eine Veränderung seiner Arbeitspraxis. Nach seiner Darstellung kontrolliert der Mittelbau inzwischen vor allem automatisch erzeugte Codierungen, und für Publikationen stehen statt Monaten zunehmend nur noch Wochen zur Verfügung, auch wegen des Drucks, weitere Fördermittel einzuwerben. Die Schilderung illustriert für mich eine mögliche Veränderung typischer Arbeitspraktiken sehr anschaulich. Es geht nicht nur darum, ob ein einzelner Button nützlich ist. Es geht auch um Erwartungen daran, wie schnell Forschung künftig erledigt werden soll.

3

TEIL 3

Aktuelle Ansätze

Methodisierte Integrations- und Innovationsvorschläge

Was die Vorschläge leisten und was wir mitnehmen können

Methodisierte Integrations- und Innovationsvorschläge

Jetzt komme ich zu Vorschlägen, die versuchen, die technischen Möglichkeiten und die methodischen Leitlinien miteinander zu verzahnen. Wie können wir Large Language Models in qualitative Forschungsmethodik einbeziehen und dabei an den Prinzipien anknüpfen, die wir für diese Forschung entwickelt haben? Ich unterscheide dabei zwischen Integrationsvorschlägen und Innovationsvorschlägen. Bei der Integration wird gefragt, wie sich LLMs in bestehende Verfahren qualitativer Textanalyse einbinden lassen. Andere Vorschläge entwickeln eine veränderte Verfahrenslogik. Ich stelle Ihnen eine Auswahl vor und erzähle jeweils auch, was ich selbst daraus mitgenommen habe. Zwei dieser Vorschläge haben wir mitentwickelt.

Die dokumentarische Methode als Prompt

Ein früher Vorschlag stammt von Burkhard Schäffer und Fabio Roman Lieder. Sie haben 2024 beschrieben, wie sich die dokumentarische Methode mit Large Language Models verbinden ließe. Ihr Ansatz heißt Reconstructive Social Research Prompting, kurz RSRP. (Lieder / Schäffer 2024)

Mein Gewinn aus diesem Artikel war zunächst ein anderes Verständnis davon, was in einen Prompt hineingehört. Ein Satz wie „Mach mir mal eine rekonstruktive Analyse“ lässt sehr viele Entscheidungen offen. Schäffer und Lieder haben verschiedene Bestandteile herausgearbeitet, die diese Entscheidungen präziser machen. Ich habe ihre Promptarchitektur für den Vortrag in fünf Bereiche zusammengefasst.

Erstens gehören Projektinformationen hinein: Was ist das für ein Forschungsprojekt? Welche Forschungsfrage haben wir? Welches Material verwenden wir, und was ist über die Untersuchungsgruppe und die Datenerhebung zu wissen? Das ist der Research-Project-Anteil.

Dann kommen grundlegende theoretische Annahmen hinzu. Wie glauben wir eigentlich, dass Wissen entsteht? Welche Vorstellungen liegen unserer Arbeit und dem Zusammenspiel von

Menschen und technischen Hilfsmitteln zugrunde? Es macht einen Unterschied, aus welcher wissenschaftlichen Tradition heraus wir auf das Material schauen. Solche Annahmen gehören in den Basic-Theory-Anteil.

Der Object-Theory-Anteil betrifft den konkreten Gegenstand. Was hat die Forschung zu diesem Thema bisher herausgefunden? Welche zentralen Konzepte gibt es, und was ist an der spezifischen Forschungssituation besonders? Auch das muss das Modell nicht erraten, sondern wir geben es mit.

Erst dann kommen wir zu dem Anteil, in dem die spezifische methodologische Perspektive der dokumentarischen Methode deutlich wird. Die dokumentarische Methode schaut nicht nur auf das, was offensichtlich in den Wörtern steht. Sie versucht auch, implizites Wissen durch die Art der Kommunikation zu erschließen. Was heißt das, und wie soll dabei vorgegangen werden? Wir erklären dem LLM die besondere Art und Weise, in der wir mit dem Material arbeiten wollen.

Ganz am Ende kommen die konkreten Analyseschritte hinzu: beispielsweise die Segmentierung des Materials, die formulierende Interpretation und die reflektierende Interpretation. Auch da reicht der Name eines Arbeitsschrittes allein nicht unbedingt aus. Was macht denn ein Mensch, der eine formulierende Interpretation erstellt? Was soll mit der Passage geschehen? Das muss in der Instruktion deutlich werden, ähnlich wie wir es in einer Seminarsitzung erklären würden.

Sie sehen, dass ein solcher Prompt umfangreich werden kann. In dem veröffentlichten Beispiel folgen Projektinformationen, theoretische Annahmen, methodologische Vorgaben und Arbeitsanweisungen aufeinander. Das ist alles noch ohne das Datenmaterial; das kommt anschließend dazu. Die Länge ist dabei kein Gütesiegel für sich. Entscheidend ist, dass die für diese Aufgabe relevanten Entscheidungen sichtbar werden.

Als ich das gelesen habe, dachte ich: Okay, jetzt verstehe ich, warum ein Prompt so lang werden kann und welche Idee dahintersteckt. Wir haben das anschließend für eine andere Methode ausprobiert. Die Ausgabe sah dieser methodischen Vorgehensweise schon sehr viel ähnlicher als bei einer kurzen Aufforderung, eine rekonstruktive Analyse zu machen.

Das finde ich ein wichtiges Verständnis: Meine Instruktionen haben einen Einfluss auf die Ausgabe, und diesen Einfluss kann ich mitgestalten. Und lesbare Prompts kann ich zudem reflektieren. Dann kann ich beispielsweise fragen: Warum fehlt hier ein bestimmter Arbeitsschritt? Hat er nicht funktioniert, wurde er bewusst weggelassen, oder müsste ich ihn ergänzen? Diese Auseinandersetzung wird möglich, wenn die Prompts zugänglich sind.

Qualitative Inhaltsanalyse mit vielen LLM-Teilschritten

Ein zweiter Vorschlag stammt von Udo Kuckartz und Stefan Rädiker. Sie haben in der sechsten Auflage ihres Buches zur qualitativen Inhaltsanalyse ein Kapitel zur Unterstützung durch KI aufgenommen. Das Kapitel ist auch gesondert zugänglich. (Kuckartz / Rädiker 2024)

Mein Gewinn daraus war die Zerlegung der Aufgabe. Die Inhaltsanalyse ist ja der Oberbegriff für viele verschiedene Arbeitsschritte. Wenn ich einem LLM nur schreibe: „Mach eine Inhaltsanalyse", habe ich diese Schritte noch nicht näher bestimmt. Kuckartz und Rädiker unterscheiden dreizehn Teilschritte und geben dafür beispielhafte Prompts.

Grob lassen sich fünf Bereiche festmachen. Zunächst geht es um die Datenerkundung, dann um die Kategorienentwicklung, das Codieren, die Analyse der codierten Daten und schließlich die Arbeit am Ergebnisbericht. Innerhalb dieser Bereiche gibt es verschiedene Unteraufgaben. Dazu gehören etwa prägnante Zusammenfassungen von Interviews, eine interaktive Exploration durch gezielte Fragen an das Material oder die Suche nach thematisch verwandten Aussagen.

Bei der Kategorienentwicklung können Vorschläge zur Systematisierung offener Codierungen und zur Ausdifferenzierung bestehender Kategorien entstehen. Beim Codieren geht es unter anderem um die Zuordnung zu bereits vorhandenen Kategorien. Die Analyse der codierten Daten kann fall- und themenbezogene Vergleiche oder die Identifikation von Beziehungsmustern zwischen Kategorien umfassen. Schließlich kann KI bei der Strukturierung der Argumentation und der sprachlichen Bearbeitung eines Berichtes eingesetzt werden.

Das sind sinnvolle Teilaspekte. Die Zerlegung macht deutlich, dass wir nicht alles auf einmal erledigen lassen müssen, sondern verschiedene Aufgaben mit jeweils eigenen Anforderungen betrachten können.

Die beispielhaften Prompts sind allerdings häufig nur ein oder zwei Zeilen lang. Da hätte ich mir mehr Ausgestaltung gewünscht. Wir haben am Zusammenfassungsbeispiel gesehen, wie viel mit einer kurzen Formulierung offenbleibt. „Mach eine prägnante Zusammenfassung" kann zu einer anderen Ausgabe führen als eine Instruktion, in der ich meine Forschungsfrage und die für mich wichtigen Aspekte genauer bestimme. Diese genauere Ausgestaltung passt vielleicht besser zu meinem Vorhaben.

Fragen an das Material und das LLM als Sparringspartner

Zwei eng verwandte Vorschläge stammen von David Morgan und Susanne Friese. Bei beiden geht es um eine veränderte Verfahrenslogik: Wir treten in einen direkten, fragenden und dialogischen Austausch mit dem Modell über das Datenmaterial, statt zunächst über die klassische Codierung zu gehen. Morgan nennt seinen Ansatz Query-Based Analysis (Morgan 2026)

Warum haben wir überhaupt codiert? Ein Grund ist, dass uns ein strukturiertes Codesystem hilft, bei vielen Daten den Überblick zu behalten. Wir sortieren Textstellen gewissermaßen in verschiedene Pappschachteln und können uns dann einem Thema nach dem anderen zuwenden. Wir müssen nicht jedes Mal das gesamte Datenmaterial durchsehen.

Was wollen wir damit aber letztlich tun? Wir wollen unsere Forschungsfrage und die damit verbundenen Detailfragen mit dem Material beantworten. Morgan schlägt vor, dafür KI zu verwenden: Wir stellen Fragen, und das LLM erzeugt dazu Antworten auf Basis des Materials. Daher auch der Name Query-Based Analysis, also eine fragenbasierte Analyse.

Die Prozessschritte lassen sich in drei Bereiche unterteilen. Zuerst gibt es breite, initiale Fragen. Danach folgen spezifischere Folgefragen. Schließlich geht es um die Suche nach unterstützenden Daten.

Positiv finde ich, dass Morgan innerhalb dieser Bereiche menschliche und KI-gestützte Arbeitsschritte unterscheidet. Der Mensch etabliert zunächst den Kontext, formuliert erste Fragen und entwickelt die entsprechenden Prompts. Die KI verarbeitet Text, identifiziert

Muster und schlägt Themen vor. Dann ist der Mensch wieder gefragt und bewertet die ersten Antworten kritisch. Darauf können präzisere Vertiefungsfragen folgen, anschließend weitere Ausgaben und wiederum eine menschliche Verdichtung und Steuerung des Dialogs. Auch die Prüfung von Textbelegen und ihrer Aussagekraft bleibt beim Menschen.

Wir haben also einen iterativen Wechsel. Und schon die Formulierung der ersten Frage ist voraussetzungsreich: Wie muss ich sie gestalten, damit eine für mein Vorhaben brauchbare Antwort entsteht? Das ähnelt der Frage, wie ich einen Prompt für eine passende Zusammenfassung entwickle.

In eine ganz ähnliche Richtung weist der Vorschlag von Susanne Friese. Sie nennt ihren Ansatz Conversational Analysis to the Power of AI und beschreibt ebenfalls einen dialogischen Austausch mit dem Modell über das Datenmaterial (Friese 2026)

Grob gesagt nutze ich das LLM dabei wie eine mitarbeitende Figur in einem Sparringteam. Ich frage: Was halten Sie davon? Was könnten wir an dieser Stelle sehen? Ich habe dazu noch eine andere Idee. Darauf reagiert das Modell, und ich arbeite mit seiner Reaktion weiter. Die Auswertung entwickelt sich in diesem iterativen Austausch. In beiden Ansätzen rückt damit das auf das Material bezogene Gespräch mit dem Modell an die Stelle des bisherigen Codierens.

Für mich steckt in solchen dialogischen und fragenbasierten Zugängen eine besonders große Herausforderung in einer kleinen Formulierung: Für die kritische Bewertung der Antworten sei Datenfamiliarität nötig. Was heißt das eigentlich? Ich muss beurteilen können, ob die Darstellung dem Material angemessen ist. Wenn etwas offensichtlich unsinnig formuliert ist, fällt mir das vielleicht sofort auf. Aber was darin nicht steht, sehe ich an der Ausgabe gerade nicht. Das Tutoriumsbeispiel hat gezeigt, warum die Kontrolle der angeführten Belege allein nicht reicht. Wie erarbeite ich mir diese Vertrautheit? Bei einer kleinen Passage kann ich das Material wiederholt lesen und mich intensiv damit auseinandersetzen. Bei zehn Stunden Interviews haben wir schnell 150 bis 200 Seiten Text. Wenn ich diese Menge einmal durchgelesen habe, kenne ich damit noch nicht alle Zusammenhänge, Gegenstellen und Unterschiede, die für die Beurteilung einer übergreifenden Darstellung wichtig sein können.

Dabei müssen wir differenzieren: Eine eng begrenzte Frage an wenige, bereits gut bekannte Passagen stellt überschaubare Prüfanforderungen. Wenn es aber um eine umfassende Synthese über viele Interviews hinweg geht, muss ich das Material in seiner gesamten Breite erschließen, Fälle vergleichen sowie Unterschiede und Zusammenhänge selbst herausarbeiten. Nach meiner Einschätzung nähern wir uns bei solchen übergreifenden Auswertungen wieder dem Aufwand einer eigenen, manuellen Analyse an – und genau dieser Unterschied verschwindet leicht hinter dem kurzen Hinweis „Datenfamiliarität nötig".

Das finde ich herausfordernd. Mit zunehmender Materialmenge habe ich deshalb Bauchweh dabei, einfach zu sagen: Lassen Sie das LLM die Antworten erzeugen, und prüfen Sie sie anschließend. Wir müssen mitbedenken, auf welcher Grundlage diese Prüfung überhaupt verantwortbar möglich ist.

Hybrides Interpretieren mit drei dialogisch integrierten LLMs

Jetzt komme ich zum ersten Vorschlag, einen, den wir selbst entwickelt haben: dem hybriden Interpretieren mit drei unterschiedlichen, dialogisch integrierten Large Language

Models. „Wir" sind Uwe Krähnke, Thorsten Pehl und ich. Unser Preprint erschien Anfang 2025. (Krähnke / Pehl / Dresing 2025), die Druckfassung im Oktober 2026. Ich will Ihnen die Idee und ihre Herleitung erzählen. Wir haben uns an ein Arbeitsformat erinnert, das in der qualitativen Forschung seit Jahrzehnten etabliert ist: die Interpretationswerkstatt.

In einer Interpretationswerkstatt diskutieren Menschen mit verschiedenen wissenschaftlichen und persönlichen Erfahrungshintergründen gemeinsam über dargebotenes Datenmaterial. Eine Person kommt vielleicht aus der Systemtheorie, eine andere aus der Psychoanalyse und eine dritte aus den Naturwissenschaften. Alle tragen ihre jeweilige Brille und damit auch unterschiedliche Perspektiven an das Material heran. Gerade daraus kann in der gemeinsamen Arbeit etwas Neues entstehen.

Die Person, die ihr Material mitbringt, gestaltet den Prozess wesentlich mit. Sie wählt einen Ausschnitt aus, formuliert eine erste Position und bringt Fragen mit. Meistens geht es um kleine Materialmengen, vielleicht eine, zwei oder drei Seiten. Sie hat diese Seiten vorher gelesen und überlegt, worüber sie mit den anderen sprechen möchte. Dann wird diskutiert. Die erste Person sagt etwas. Die zweite versteht das nicht und fragt zurück. Eine dritte hat noch eine andere Idee. So können zehn oder zwanzig Menschen anderthalb Stunden über wenige Seiten Text sprechen und versuchen zu verstehen, wie das Gelesene gedeutet werden kann.

Dieser Weg lässt sich dokumentieren. Wir können die Sitzung aufnehmen und transkribieren. Wenn später jemand fragt: Wie sind Sie auf diese Deutung gekommen?, können wir zeigen: Hier war die erste Idee. Die haben wir aus diesen Gründen abgelehnt. Etwas anderes haben wir zunächst so verstanden, dann aber unter Einbezug weiterer Argumente anders gelesen. Wir können nachvollziehbar machen, wie wir von Hütchen auf Stöckchen gekommen sind.

Dabei kann eine andere Person in der Werkstatt eine Interpretationsidee haben, auf die ich selbst nicht gekommen wäre. Vielleicht sitzt dort eine Person, die etwas überzeugend herausarbeitet. Ich greife den Gedanken auf und verwende ihn in meiner weiteren Arbeit. Die Erkenntnis ist im gemeinsamen Austausch entstanden; ich muss mich zu ihr verhalten und begründen können, weshalb ich sie aufnehme, darf sie aber in meiner Arbeit zentral aufgreifen, so der bisherige commen-sense.

Unsere Idee ist nun, eine solche Arbeitsanordnung mit Large Language Models nachzubilden. Die forschende Person bleibt die steuernde Instanz. Sie selektiert ein kleines Datenstück und schreibt zunächst eine eigene Erstinterpretation. Dieses Material und die Erstinterpretation bekommt das erste LLM zusammen mit einem Initialprompt. Die dafür entwickelten Prompts haben wir offen zugänglich aufbereitet. Das Modell soll sich in seiner Antwort auf Material und Deutung beziehen und einen Beitrag liefern, mit dem wir weiterarbeiten können: „Wenn ich auf diese Stelle schaue, sehe ich das so und so." Mit dieser Antwort gehe ich zum zweiten LLM. Es erhält das Originalmaterial, meine Erstinterpretation und den Beitrag des ersten Modells. Dadurch kann es sich sowohl auf mich und die Passage als auch auf die erste Modellausgabe beziehen. Und die Modelle können sich durchaus widersprechen. Ein zweites Modell sagt vielleicht sinngemäß: „Das sehe ich überhaupt nicht so. Man könnte die Stelle auch ganz anders verstehen." Dadurch wird sichtbar, dass die erste Ausgabe keine feststehende Wahrheit ist. Unterschiedliche Perspektiven auf denselben Text können entstehen. Dann gebe ich den bisherigen Verlauf an ein drittes LLM. Dieses bekommt ebenfalls das Material, meine Erstinterpretation und die Beiträge der beiden anderen Modelle. Es kann sich auf alle diese Texte beziehen und eine

weitere Deutung anbieten. Jetzt bin ich wieder gefragt. Ich lese die Beiträge und positioniere mich dazu. Vielleicht sage ich: Das verstehe ich nicht, bitte erkläre das genauer. Oder: Dieser Gedanke ist spannend, aber nur diese eine Facette. Den Rest finde ich nicht materialnah oder zu weitgehend. Lasst uns diese Stelle vertiefen.

Ich moderiere und arbeite an der Interpretation entscheidend mit. Ich entscheide, welche Richtung weiterverfolgt wird und was ich zurückweise. Mit meiner Zwischenposition beginnt die nächste Runde. Das erste Modell bekommt nun auch die Beiträge der anderen Modelle und meine Reaktion darauf. Durch die Weitergabe des Gesprächsverlaufs haben die Modelle jeweils Zugriff auf den bis dahin entstandenen Austausch.

Nach drei oder vier solchen Runden haben wir Deutungen erarbeitet, die ich selbst intensiv mit durchdrungen habe. In unseren bisherigen Erprobungen eröffnete dieses Vorgehen eine größere Perspektivenvielfalt und führte zu differenzierteren Lesarten als einzelne, unmoderierte LLM-Anfragen zum selben Material, auch wenn ein systematischer empirischer Vergleich noch aussteht. Der Verlauf dokumentiert, welche Vorschläge aufkamen, wie ich mich dazu verhalten habe und was daraus entstanden ist. Diese Wechselwirkung zwischen Modellen und forschender Person ist die Grundidee des hybriden Interpretierens.

Damit greifen wir einige der vorhin genannten Herausforderungen auf. Bias und Varianz werden als mögliche Quellen unterschiedlicher Lesarten in einen moderierten Prozess eingebracht. Die Unterschiede können eine produktive Spannung erzeugen. Die Materialauswahl, die Erstdeutung, die Zwischenpositionierungen und die abschließende Synthese bleiben Aufgaben der forschenden Person.

Mehrere Modelle garantieren dabei keine angemessene Interpretation. Sie reagieren aufeinander und können auch denselben Gedanken übernehmen. Die Beiträge müssen weiterhin am Material beurteilt werden. Das dokumentierte Gespräch macht jedoch den Weg unserer Auseinandersetzung transparent: Dort lässt sich im Einzelnen nachvollziehen, mit welchen Argumenten ich mich für oder gegen eine bestimmte Lesart entschieden habe und worin mein eigener Beitrag zur Interpretation bestand.

Umsetzen lässt sich das mit lokalen Modellen oder geeigneten Chatoberflächen der Hochschulinfrastruktur, etwa der Academic Cloud. Auch in der f4 2026 Softwareoberfläche haben wir diesen Ablauf kostenfrei ermöglicht. Das Verfahren hängt also nicht an einem einzigen Zugang. Zum Nacharbeiten finden Sie den Preprint zum hybriden Interpretieren, ein Praxisbeispiel und eine Einführung mit den verwendeten Prompts.

Serendipity-Prompting als Anstoß und Irritation der eigenen Interpretationsarbeit

Unsere zweite Idee ist das Serendipity-Prompting. Dabei haben wir versucht, die Zielrichtung der Modellnutzung zu drehen. Wir wollen weg von der Nutzung im Stil einer Frage-Antwort-Dialektik hin dazu, dass uns LLM-Ausgaben irritieren, widerständig zu unseren Positionen sind und Fragen eröffnen, die uns zurück zum Material führen. Der Gedanke ist an ein sokratisches, mäeutisches Prinzip angelehnt. (Krähnke / Dresing / Pehl 2026)

Damit das Modell auf unsere Perspektive reagieren kann, muss es wissen, wo wir stehen. Deshalb wählen wir zuerst eine kleine Passage aus und formulieren eine eigene Erstdeutung.

Erst auf dieser Grundlage erzeugt das LLM drei verschiedene Impulse. Zunächst kommt ein Validierungsimpuls. Er greift unsere Lesart auf: Was spricht dafür? Gibt es in der Passage eine weitere Stelle, die diese Deutung stützen könnte? Danach kommt ein Falsifikationsimpuls, also eine Gegenlesart: Moment mal, man könnte das Phänomen auch anders deuten. Beide Angebote sollen am Material argumentieren. Damit ist eine Spannung entstanden: Es könnte so sein, aber auch anders. Am Ende steht ein Syntheseimpuls mit offenen Fragen. Was lässt sich jetzt weiterverfolgen? Wo müsste ich genauer hinschauen oder weiterlesen? Die vorläufigen Deutungsangebote sollen meine eigene Auseinandersetzung anregen.

Ich zeige das an der fiktiven Lehrpassage aus den Folien. Wir kommen dafür zu der Stelle zurück, in der die befragte Person „sie" sagt. Nun nehmen wir noch eine weitere Frage und Antwort dazu:

> ***I:*** *Zum Anfang: Wie lange sind Sie genau schon bei move35 hier in Marburg dabei?*
>
> ***B:*** *Das erste Treffen hat im September stattgefunden. Das war '23. Und seit diesem Punkt sind sie auch ganz aktiv.*
>
> ***I:*** *Und wie würden Sie Ihre Arbeit beschreiben?*
>
> ***B:*** *Meine Arbeit ist eigentlich die Koordination dieses ganzen Treffens. Und der Versuch, es voranzubringen.*

Meine erste Deutung dazu lautet:

> *An dem „sie" bleibe ich hängen. Es klingt distanziert. Ich glaube, da gibt es Probleme mit der Gruppe.*

Darauf reagiert das Modell in dem dokumentierten Beispiel mit drei Impulsen:

> ***Validierungsimpuls:*** *Du bleibst an dem „sie" in „sind sie auch ganz aktiv" hängen. Es klingt für dich distanziert, und du vermutest dahinter Probleme mit der Gruppe. Dabei fällt etwas auf, das schon vor dem „sie" beginnt und das deine Deutung noch nicht nennt. Die Einstiegsfrage richtet sich an B. persönlich, doch die Antwort beginnt mit einem Ereignis, „Das erste Treffen". Genau da, wo nach der eigenen Beteiligung gefragt ist, spricht B. nicht von sich. Das kann man als Distanz lesen.*
>
> ***Falsifikationsimpuls:*** *Bei der zweiten Frage spricht B. von sich selbst: „Meine Arbeit ist". Das „sie" aus der ersten Antwort kann man auch als Gegenstück zu diesem „Meine" lesen. Auf der einen Seite das, was „sie" tun, auf der anderen die eigene Aufgabe. Der Wechsel von „sie" zu „Meine Arbeit" kann als Rollenunterscheidung gelesen werden und muss nicht Distanz bedeuten.*
>
> ***Syntheseimpuls:*** *Deine Lesart lässt noch offen, was das „sie" über B.s Verhältnis zur Gruppe sagt. Dass es Probleme gibt, ist bisher deine Vermutung. Wie passen für dich das „sie" aus der ersten und das „Meine" aus der zweiten Antwort zusammen, und was ändert sich dadurch an deiner Vermutung?*

Sie sehen: Die Frage geht an mich zurück. Ich muss überlegen, ob ich die Gegenlesart plausibel finde und wie ich meine eigene Vermutung im Verhältnis zu der erweiterten Passage bewerte.

Das ist unsere Idee. Sie zielt auf Inspiration für eine intensive Auseinandersetzung mit dem Material. Wir arbeiten hier vielleicht mit fünf oder zehn Zeilen, einem Absatz oder einer kleinen Passage. Bei diesem Umfang können wir die Vorschläge unmittelbar am Text prüfen und weiterdenken. Das konterkariert allerdings die Vorstellung, mit KI vor allem schneller fertig zu werden.

Der mögliche Gewinn liegt in einer Perspektive, die wir vorher nicht gesehen haben, oder in einer Frage, die wir an unsere eigene Deutung noch nicht gestellt hatten.

Die Beschreibung des Ansatzes finden Sie in unserem Manifest zum Serendipity-Prompting und zur methodisierten Neugier.

Was die Vorschläge leisten und was wir mitnehmen können

Das waren ausgewählte Vorschläge, bestehende qualitative Prinzipien und Verfahrensideen mit LLMs zu verbinden. Es sind Vorschläge in einem sich schnell verändernden Feld. Damit ist noch nicht geklärt, wie gut sich welches Vorgehen unter welchen Bedingungen bewährt, wie es vermittelt werden kann und welche Formen der Prüfung erforderlich sind.

Die umfassende Erprobung solcher Verfahren braucht Zeit. Man müsste Erfahrungen aus verschiedenen Projekten zusammentragen, die Bedingungen der Anwendung genauer bestimmen und die Folgen für den Forschungsprozess untersuchen. Das passt nur begrenzt zum Tempo, in dem neue technische Angebote erscheinen.

Hinzu kommt, dass sich mit den technischen Möglichkeiten auch Erwartungen im Wissenschaftsbetrieb und in der Forschungsförderung verschieben. Wir arbeiten bei audiotranskription seit 22 Jahren zu wissenschaftlicher Transkription. Früher hatten wir bei zeitweise vierzig Schreibkräfte beschäftigt und für Forschungsprojekte im Vier-Augen-Prinzip transkribiert; Budgets von zehntausend Euro waren dafür keine ungewöhnliche Größenordnung. Mit der automatischen Transkription werden für diese Tätigkeit heute deutlich weniger Mittel eingeplant und Schreibkräfte haben wir schon lange nicht mehr. Wann kommt ein ähnlicher Gedanke bei der Auswertung auf? Wann fragt eine fördernde Institution: Brauchen wir dafür mehrere wissenschaftliche Mitarbeitende für mehrere Jahre, oder reicht eine Softwarelizenz mit KI-Funktionen?

Wir stehen deshalb vor Entscheidungen: Welche Arbeit möchten wir übertragen? Was müssen wir selbst leisten, um das Ergebnis verantworten zu können? Welche Konsequenzen ziehen wir aus den Güteanforderungen, denen wir uns verpflichtet fühlen? Und welche Formen qualitativer Datenanalyse sollten wir noch selbst lernen oder vermitteln, wenn scheinbar einige automatisierbar sind?

Meine Sorge ist, dass diese Entscheidungen vor allem durch die verfügbaren Werkzeuge und den alltäglichen Arbeitsdruck vorweggenommen werden. „Guck mal, es funktioniert" könnte dann an die Stelle einer Auseinandersetzung darüber treten, was genau funktioniert und ob es für unsere Forschung angemessen ist. Ich glaube, dabei könnten wir etwas verlieren.

Für viele der angesprochenen Herausforderungen haben wir noch keine abschließende Lösung, während die Funktionen gleichzeitig bereits genutzt werden. Umso wichtiger ist die eigene praktische und kritische Auseinandersetzung: Eigene Erfahrungen helfen, Angebote realistisch einzuschätzen, Positionierungen zu schärfen und die Argumente anderer

nachzuvollziehen. Beteiligen Sie sich, entwickeln Sie Ideen, Werkzeuge und begründete Umgangsweisen.

Vielen Dank für Ihre Aufmerksamkeit und die Zeit, die Sie sich genommen haben. Bleiben Sie wachsam, behalten Sie die gute Laune, und gestalten Sie die Veränderungen mit.

Anhang

Quellen und Materialien zum Nacharbeiten

Die folgenden Nachweise führen zu den besprochenen Studien, methodischen Vorschlägen und Werkzeugen. Preprints sind als solche gekennzeichnet. Bei Büchern führen die Links zur Verlagsseite; soweit ein offen zugänglicher Text vorliegt, ist er zusätzlich verlinkt. Der technische Stand und die beschriebenen Angebote beziehen sich auf November 2026.

Studien und Beispiele

Li, Tianshi (2026). Agentic LLMs as Powerful Deanonymizers: Re-identification of Participants in the Anthropic Interviewer Dataset. Preprint. Quelle öffnen.

Laudel, Grit; Bielick, Jana (2019). Forschungspraktische Probleme bei der Archivierung von leitfadengestützten Interviews. Forum Qualitative Sozialforschung / Forum: Qualitative Social Research, 20(2), Art. 10. Quelle öffnen. Weiterer Zugang. Das Qualiservice-Beispiel steht in Absatz 55.

Mozygemba, Kati; Hollstein, Betina (2023). Anonymisierung und Pseudonymisierung qualitativer textbasierter Forschungsdaten: eine Handreichung. Quelle öffnen.

Shaw, Steven D.; Nave, Gideon (2026). Thinking–Fast, Slow, and Artificial: How AI Is Reshaping Human Reasoning and the Rise of Cognitive Surrender. Preprint. Quelle öffnen.

Lee, Hao-Ping; Sarkar, Advait; Tankelevitch, Lev; Drosos, Ian; Rintel, Sean; Banks, Richard; Wilson, Nicholas (2025). The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers. Quelle öffnen. Weiterer Zugang.

Kosmyna, Nataliya; Hauptmann, Eugene; Yuan, Ye Tong; Situ, Jessica; Liao, Xian-Hao; Beresnitzky, Ashly Vivian; Braunstein, Iris; Maes, Pattie (2025). Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task. Preprint. Quelle öffnen.

Budzyń, Krzysztof, et al. (2025). Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study. The Lancet Gastroenterology & Hepatology, 10(10), 896–903. Quelle öffnen. Weiterer Zugang.

Gerlich, Michael (2025). From Offloading to Engagement: An Experimental Study on Structured Prompting and Critical Reasoning with Generative AI. Data, 10(11), 172. Quelle öffnen. Weiterer Zugang.

Hughes, Kahryn; Tarrant, Anna (2026). Is AI influencing qualitative data histories and futures? Blogbeitrag. Quelle öffnen.

Methoden und Fachliteratur

Dresing, Thorsten; Pehl, Thorsten; Krähnke, Uwe (2026). Datenmaskierung mittels Pseudo- und Anonymisierung reicht nicht! Methodische, ethische und datenschutzrechtliche Kriterien bei der KI-gestützten Verarbeitung qualitativer Daten. Preprint. Quelle öffnen.

Dresing, Thorsten; Krähnke, Uwe; Hense, Jan; Pehl, Thorsten (2026). Pseudo- und Anonymisierung reichen nicht! Methodische, ethische und datenschutzrechtliche Herausforderungen bei der KI-gestützten Verarbeitung personenbezogener qualitativer Daten. Zeitschrift für Evaluation, 25(2), 394–407. Quelle öffnen.

Lieder, Fabio Roman; Schäffer, Burkhard (2024). Reconstructive Social Research Prompting (RSRP). Distributed Interpretation between AI and Researchers in Qualitative Research. Preprint. Quelle öffnen.

Lieder, Fabio Roman; Schäffer, Burkhard (2026). Reconstructive Social Research Prompting. Distributed Interpretation Between AI and Researchers. In Digital Hermeneutics II, S. 235–249. Springer. Quelle öffnen.

Schäffer, Burkhard; Lieder, Fabio Roman; Krämer, Franz (2026). Interpretieren zusammen mit KI. Einführung in Theorie und Praxis qualitativer Forschung mit generativen Sprachmodellen. Springer VS, Wiesbaden. Quelle öffnen. Weiterer Zugang.

Kuckartz, Udo; Rädiker, Stefan (2024). Qualitative Inhaltsanalyse. Methoden, Praxis, Umsetzung mit Software und künstlicher Intelligenz. 6., überarbeitete und erweiterte Auflage. Beltz Juventa. Quelle öffnen. Weiterer Zugang. Vollständiges KI-Kapitel in englischer Sprache.

Morgan, David L. (2026). Query-Based Analysis: A Strategy for Analyzing Qualitative Data Using ChatGPT. Qualitative Health Research, 36(2–3), 206–217. Quelle öffnen. Weiterer Zugang.

Friese, Susanne (2026). From Coding to Conversation: A New Methodological Framework for AI-Assisted Qualitative Analysis. Qualitative Inquiry, OnlineFirst. Quelle öffnen. Weiterer Zugang.

Krähnke, Uwe; Pehl, Thorsten; Dresing, Thorsten (2025). Hybride Interpretation textbasierter Daten mit dialogisch integrierten LLMs: Zur Nutzung generativer KI in der qualitativen Forschung. Preprint. Quelle öffnen.

Krähnke, Uwe; Dresing, Thorsten; Pehl, Thorsten (2026). Hybrid Interpretation of Text-Based Data with Dialogically Integrated LLMs. On the Use of Generative AI in Qualitative Research. In S. Friese & D. L. Morgan (Hrsg.), Qualitative Data Analysis With AI: Theory, Methods, and Practice. Sage. Quelle öffnen.

Krähnke, Uwe; Dresing, Thorsten; Pehl, Thorsten (2026). Serendipity-Prompting und methodisierte Neugier: Manifest für eine mäeutische Nutzung generativer KI in der qualitativen Forschung. Preprint. Quelle öffnen.

Friese, Susanne; Morgan, David L. (Hrsg.) (2026). Qualitative Data Analysis With AI: Theory, Methods, and Practice. Sage. Quelle öffnen.

Kempny, Christian; Annac, Kübra; Yilmaz-Aslan, Yüce; Brzoska, Patrick (2026). KI in der qualitativen Forschung: Von der Studienplanung bis zur Datenauswertung. Springer, Berlin und Heidelberg. Quelle öffnen.

von Unger, Hella; Jeanrenaud, Yves; Raucheger, Thomas (Hrsg.) (2026). Qualitativ Forschen mit KI über KI: Ein soziologisches Lehrforschungsprojekt. Springer VS, Wiesbaden. Quelle öffnen.

Jowsey, Tanisha; Braun, Virginia; Clarke, Victoria; Lupton, Deborah; Fine, Michelle (2025). We Reject the Use of Generative Artificial Intelligence for Reflexive Qualitative Research. Qualitative Inquiry, OnlineFirst. Quelle öffnen.

Kuckartz, Udo; Dresing, Thorsten; Rädiker, Stefan; Stefer, Claus (2008). Qualitative Evaluation. Der Einstieg in die Praxis. 2. Auflage. VS Verlag für Sozialwissenschaften, Wiesbaden. Quelle öffnen.

Rädiker, Stefan (2026). AI-Assisted Analysis of Qualitative Data with MAXQDA: Practical Guidance and Methodological Reflections. Berlin: MAXQDA Press. Quelle öffnen. Weiterer Zugang.

Belkin, Sam; White, Jacob; Burke, Cale G.; Cepek, Emma (2026). Transforming qualitative research: The AQUATIC approach to AI-driven data analysis. DIGITAL HEALTH, 12, 20552076261435085. Quelle öffnen. Weiterer Zugang.

Werkzeuge und technische Hintergründe

OpenAI (2026). Sharing AI progress in mathematics. Quelle öffnen.

VERBI Software (2026). MAXQDA Update 27.0: Local AI, smarter code grouping, and more. Quelle öffnen.

Element Labs (o. J.). LM Studio. Quelle öffnen.

Hugging Face (o. J.). Models – Hugging Face. Quelle öffnen.

VERBI Software (o. J.). MAXQDA Tailwind. Quelle öffnen.

Liu, Nelson F.; Lin, Kevin; Hewitt, John; Paranjape, Ashwin; Bevilacqua, Michele; Petroni, Fabio; Liang, Percy (2024). Lost in the Middle: How Language Models Use Long Contexts. Transactions of the Association for Computational Linguistics, 12, 157–173. Zeitschriftenartikel. Quelle öffnen. Weiterer Zugang.

VERBI Software (o. J.). AI Coding: Documents. Quelle öffnen.

Regelungen und wissenschaftliche Dokumentation

Europäisches Parlament und Rat der Europäischen Union (2016). Verordnung (EU) 2016/679 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten. Rechtsquelle. Quelle öffnen. Konsolidierter deutscher Text. Besonders relevant sind Art. 2, 6, 7, 9, 13, 28, 32, 35 und 44 ff.

Verwaltungsgericht Kassel (2026). Umgang mit Künstlicher Intelligenz bei studentischen Prüfungsleistungen. Quelle öffnen.

DeGEval – Gesellschaft für Evaluation e. V. (2017). Standards für Evaluation. Erste Revision 2016. Mainz: DeGEval. Quelle öffnen. Besonders G6, G7 und G8; verabschiedete Revision 2016, Druckausgabe 2017.

Harvard Library (o. J.). Acknowledging and Citing AI – Artificial Intelligence for Research and Scholarship. Quelle öffnen.

Deutsche Forschungsgemeinschaft (2019). Leitlinien zur Sicherung guter wissenschaftlicher Praxis. Kodex. Quelle öffnen.

Offene Anleitungen zu den eigenen Ansätzen

Einführung in das hybride Interpretieren mit drei LLMs und den verwendeten Prompts.

Praxisbeispiel zur hybriden Interpretation. Die move35-Passage ist ein fiktives Lehrbeispiel.

Publikationen und weitere Materialien von audiotranskription.