

# KI in der qualitativen Forschung?!

KI-Einsatz in der qualitativen Forschung: Zwischen Beschleunigung, Erosion und Methodeninnovation

Vortragsskript - 05. Februar 2026 – PH Tirol - Dr. Thorsten Dresing

## Einleitung

Dies ist das mithilfe von LLMs (Claude Opus 4.5, chatGPT 5.2 thinking) erzeugte Skript eines durch ASR (f4x) automatisch erzeugten Transkriptes meines 3stündigen Onlinevortrags vom 5.2.2026 an der PH Tirol. Die Inhalte wurden einmal manuell redigiert und konnten 2 Stunden nach dem Vortrag als PDF versendet werden. Es aktualisiert das erste Skript aus 10/2025.

Das Thema künstliche Intelligenz in der qualitativen Forschung löst eine bemerkenswerte emotionale Bandbreite aus – Faszination, Ablehnung oder Ambivalenz und manchmal auch alles gleichzeitig. Seit nunmehr 20 Monaten beschäftige ich mich täglich mit der Frage, was KI für qualitative Forschung bedeutet. In dieser Zeit verfolgte ich verschiedenste Pfade, habe Werkzeuge und methodische Innovationen getestet und mitentwickelt. Dieser Vortrag – in seiner 25. Durchführung seit Januar 2025 – ist dabei selbst ein Spiegel der Dynamik: Jeder Foliensatz ist anders, auch hier wurden die letzten Änderungen am Vortrag eingepflegt, weil buchstäblich wochenweise neue Hinweise, Ideen oder Studien erscheinen, die das Verständnis verändern.

Die Entwicklung vollzieht sich in atemberaubendem Tempo. Wirklich kategorial neu erscheinen mir im Januar und Februar 2026 etwa OpenClaw und Moltbook, wo 1,5 Millionen KI-Instanzen miteinander diskutieren, und neue Dinge erfinden Und Tools wie Claude Co Work. Das ist erneut ein qualitativer Sprung und deutlich anders als LLM einfach in einem Browser zu nutzen. Die Veränderungen sind nicht mehr hypothetisch und für den nächsten Vortrag werden ich das Fragezeichen im Titel streichen müssen.

Das Ziel dieses Vortrags ist nicht, eine Position vorzugeben, sondern zu informieren, zu inspirieren, gelegentlich vor den Kopf zu stoßen – und Handlungsfähigkeit herzustellen. Ich werde sowohl der ablehnenden Haltung als auch der Neugier und den methodisch Interessierten Hilfestellung und Argumente geben, ihre jeweiligen Position beizubehalten ☺ Sie als Teilnehmenden sollen in die Lage versetzt werden, danach besser informiert selbst zu entscheiden, welche Richtung sie einschlagen wollen. Denn eines ist klar: Die Transformation findet statt, ob wir mitgestalten oder nicht.

Eine Sache vorab, die mir wichtig ist: Ich bin nicht nur Forscher, der sich das aus akademischer Distanz anschaut. Ich bin Mitgründer von audiotranskription.de, wir verkaufen Transkriptions- und Analysesoftware und -dienste, und wir entwickeln gerade KI-Funktionen für unsere Software f4. Wenn ich also gleich eigene Prompts, eigene Tools

und eigene Publikationen zeige, dann tue ich das, weil es schlicht meine Arbeit ist und ich glaube, dass die Ergebnisse relevant sind – aber Sie sollten wissen, dass ich dabei nicht interessensfrei bin. Bewerten Sie das bitte entsprechend.

Ergänzend zu diesem Vortrag steht auch das alte Skript eines Vortrags aus Oktober 2025 als PDF zur Verfügung. Im Vergleich der beiden werden sie rasch erkennen können, was sich bereits in dieser kurzen Zeit an ergänzenden Details oder veränderter Bewertung zumindest für mich persönlich ergeben hat ([www.audiotranskription.de/wp-content/uploads/Skript-KI-transformiert-17102025.pdf](http://www.audiotranskription.de/wp-content/uploads/Skript-KI-transformiert-17102025.pdf) - Stand Oktober 2025, in Teilen bereits veraltet).

## Thematischer Überblick

1. Automatische Transkription funktioniert
2. Automatische Auswertung? Nutzungsstrategien, Tools & ihre Grenzen
3. Widerstände, Kontroversen und offene Probleme
4. Kann es trotzdem funktionieren? Einsichten aus einem Re-Analyse-Experiment
5. Was macht LLM-Nutzung mit uns? Empirische Befunde
6. Methodisierte Integration statt LLM/Tools allein
7. DSGVO-konforme Umsetzungspfade
8. Was kann ich jetzt tun?

## 1. Automatische Transkription: Der gelöste Fall

Es gibt einen Bereich, in dem der Einsatz von KI in der qualitativen Forschung als weitgehend unproblematisch gelten kann und auch in der Scientific Community weitgehend kritikfrei akzeptiert ist: die automatische Transkription. Seit Ende 2022 funktioniert sie zuverlässig. Allein bei audiotranskription.de haben sich über 100.000 Menschen für den Service registriert – weltweit gibt es Anbieter mit Millionen Nutzenden.

Die Zeitersparnis beträgt in vielen Fällen rund 50 Prozent gegenüber manueller Transkription. Was bleibt, ist der Korrekturdurchgang. Interessanterweise scheint dieser kognitiv weniger anspruchsvoll zu sein als manuelles Tippen, sodass freiwerdende kognitive Ressourcen produktiv genutzt werden können: Wer der qualitativen Inhaltsanalyse folgt, kann während der Korrektur bereits Fallzusammenfassungen erstellen. Anhänger der dokumentarischen Methode nehmen erste Segmentierungen vor, in der Grounded Theory lassen sich Ideen fürs offene Kodieren sammeln. Potenziert wird dies dadurch, dass man beim Korrigieren in die Originaldaten mit Reinhört und den

Wortlaut unmittelbar präsent hat. Die vorher strikt getrennten Arbeitsteile – erst Transkription, dann Analyse – verzahnen sich.

Grenzen bestehen bei Sprecherüberlappungen (Schulklassensituationen, Café-Hintergrundstimmen), bei intensiven Dialekten und bei Störgeräuschen anderer Stimmen – ein Fall zeigte, wie ein Fernseher im Nebenzimmer als zwei zusätzliche Sprecher transkribiert wurde. DSGVO-konformes Arbeiten ist möglich, sowohl über Offline-Tools wie Noscibe als auch über Anbieter mit entsprechenden Auftragsdatenverarbeitungsverträgen (sofern sie eine angemessenen Einwilligungserklärung mit den befragten unterzeichnet haben)

Die Entwicklung geht weiter: Für komplexere Regelsysteme wie das gesprächsanalytische Transkriptionsmodell GAT-2, das manuell einen Zeitaufwand von 1:20 bis 1:60 erfordert, wurde im Januar 2026 ein neuer Ansatz erprobt. Mithilfe von Claude Co Work – einer App von Anthropic, die Claude als KI-Assistenten für Softwareentwicklung nutzt – habe ich zu ersten Testzwecken in wenigen Stunden eine Software gebaut, die automatische Transkription mit einem nachgeschalteten LLM-Schritt verbindet, um komplexe Transkriptionsregeln lokal anzuwenden. Ich kann nicht programmieren; die gesamte Entwicklung erfolgte im Chatfenster mit natürlicher Sprache. Diese Entwicklung wirft eine fundamentale Frage auf: Wenn individuell zugeschnittene Software in Stunden entstehen kann, könnte sich durchaus in der mittleren Zukunft das gesamte Geschäftsmodell kommerzieller (Forschungs-)software ändern.

## 2. Nutzungsstrategien: Prompt Engineering, Vorlagen und Tools

Wenn automatische Transkription funktioniert und Forschende dabei Zeit sparen, liegt der Gedanke nahe, KI auch für die Auswertung einzusetzen. Die beobachtbaren Nutzungsstrategien im Feld lassen sich grob in drei Kategorien unterteilen: selbst entwickelte Prompt (Prompt Engineering), die Nutzung von Prompt-Vorlagen anderer sowie die Verwendung angebotener Software-Tools.

### **Erstes Standbein: Prompt Engineering**

Philipp Mayring, die prägende Figur der qualitativen Inhaltsanalyse im deutschsprachigen Raum, hat KI für eine Inhaltsanalyse ausprobiert und auf dem Berliner Methodentreffen 2024 berichtet: Das funktioniert nicht. Er tippte eine kurze Anweisung ein, gab ein Interview dazu – und erhielt Schrott. Diese Einschätzung ist nachvollziehbar: So wie er es versuchte, funktioniert es tatsächlich nicht. Aber warum?

Fünf grundlegende Probleme erklären die Problematik:

Erstens: Einfache Ein- oder Zweizeiler liefern für qualitative Forschung keine zufriedenstellenden Ergebnisse.

Zweitens: Es gibt keine Bedienungsanleitung. Selbst die Hersteller wissen nicht vollständig, wie ihre Modelle optimal zu nutzen sind. Optimale Nutzung ist Gegenstand unzähliger, laufender Forschungsprojekte.

Drittens: Die bekannten Tipps sind zahlreich – weit über 50 – und teilweise völlig kontraintuitiv. Einige sind nachvollziehbar: mehr Kontext geben, Beispiellösungen zeigen, Aufgaben in Teilschritte zerlegen (Chain of Thought). Andere wirken absurd, funktionieren aber nachweislich: Dem Modell sagen, es solle erst ein- und ausatmen. Trinkgeld anbieten. Buchstaben durch Zahlen ersetzen (E durch 3, O durch 0), sodass der Text für Menschen kaum lesbar, für das LLM aber anders verarbeitbar wird. Diese Strategien basieren auf spezifischen Eigenschaften der Trainingsdaten und der Modellarchitektur – sie sind nicht intuitiv erschließbar.

Viertens: Unterschiedliche Modelle reagieren unterschiedlich. Was bei GPT-5.2 funktioniert, scheitert oder variiert substantiell möglicherweise bei Claude 4.5. Zudem kommen neue Modellversionen mittlerweile alle zwei bis drei Monate, und jede Version verhält sich anders auf dieselben Prompts. Einmal erfolgreich entwickelte und angewendete Strategien, müssen immer wieder überprüft werden.

Fünftens: Die kompetente Nutzung erfordert Wochen an Recherche, Einarbeitung, Testzeit und Anpassung – ein Aufwand, der Alltagsmenschen mit Lehre, Prüfungen und Familie schlicht überfordert.

Es scheint sich, zumindest für uns, eine übergreifende Meta-Strategie herauszukristallisieren: Nie nur ein LLM nutzen, sondern immer iterativ mit verschiedenen Modellen arbeiten. Das erhöht die Qualität häufig massiv (aber wie bei der Sahne, zu viel/lange rühren ist nicht immer vorteilhaft).

## **Zweites Standbein: Prompt-Vorlagen**

Für diejenigen, die die Prompt-Entwicklung nicht selbst leisten können oder wollen, existieren öffentlich verfügbare Vorlagen. Die meisten guten Prompts sind entweder ziemlich umfangreich – oft zwei DIN-A4-Seiten allein für den Prompt, ohne Datenmaterial – oder implementieren eine Strategie iterativen Vorgehens. Eine kuratierte Sammlung findet sich unter: [audiotranskription.de/prompts-fuer-dein-qualitatives-forschungsprojekt/](https://audiotranskription.de/prompts-fuer-dein-qualitatives-forschungsprojekt/). Viele weitere finden sich via Recherche im Internet.

Darunter befinden sich Prompts zur Forschungsfragenfindung, Methodendesign, zur Themenschärfung für Qualifikationsarbeiten (ein Professor verlangt von Studierenden, vor der Sprechstunde iterativ mit LLMs gearbeitet zu haben), zur wissenschaftlichen Textzusammenfassung, zur Textüberprüfung sowie ein älterer Zwischenstand eines Evaluations-Prompt, der im Rahmen der nachfolgend beschriebenen Re-Analyse entwickelt wurde.

### **Drittes Standbein: Software-Tools**

Verschiedene Softwarehersteller haben KI-Funktionen in ihre Produkte integriert oder als eigenständige Dienste entwickelt. Dabei ist zu beachten: Kein kleines oder mittleres Unternehmen entwickelt eigene LLMs. Was diese Tools leisten, ist die Entwicklung spezialisierter Prompts, eingebettet in eine Programmoberfläche. Das hat Vorteile (u.a. Vereinfachung) und Nachteile (u.a. mangelnde Transparenz, fehlende Zitierbarkeit).

MAXQDA AI Assist nutzt Amazon-Bedrock-Server in der EU, auf denen Googles Gemini oder Anthropic's Claude arbeiten. Die im Hintergrund verwendeten Prompts sind für Nutzende nicht. Für wissenschaftliche Transparenz und Zitierbarkeit nach aktuellen Standards (z. B. Harvard-Empfehlungen für KI-Zitation) ist dies problematisch: Weder das genaue Modell noch die Prompts noch deren Versionsstand können angegeben werden. Die spätere Wiederholung zur Nachprüfung könnte scheitern. Ein Experiment mit der Subcode-Vorschlagsfunktion verdeutlicht sowohl Nutzen als auch Problematik: Zu einem Code mit 91 Textsegmenten wurden vier Mal hintereinander Vorschläge zur Ausdifferenzierung (Subcodes vorschlagen) generiert. Die Ergebnisse: fünf Kategorien – eine Minute später fünf andere – dann acht – dann elf. Diese Varianz ist einerseits typisch für qualitative Forschung (verschiedene Perspektiven auf dasselbe Material können gleichzeitig gültig sein). Die Idee von MAXQDA ist dennoch durchaus plausibel, nicht finales Ergebnis, sondern Vorschläge, die die eigene Perspektive bereichern sollen. Andererseits: Wie hoch ist die Wahrscheinlichkeit, dass Menschen mit wenig Methodenkompetenz und Zeitdruck solche Vorschläge eins zu eins übernehmen? Vermutlich nicht null Prozent.

MAXQDA Tailwind ist eine separate, webbasierte Lösung, die Arbeitsschritte die auch im Rahmen einer qualitativen Inhaltsanalyse benötigt werden teilautomatisiert. Nach dem Hochladen von Interviews werden automatisch Fallzusammenfassungen erstellt – mit Verlinkungen zu den Originalpassagen (grüne Zahlen) zur Überprüfung. Ein Topic-Reiter identifiziert induktiv Themen mit Operationalisierung, generiert deskriptive Ergebnisberichte und der Chatassistent kann bspw. auf Anfrage Vorschläge für eine Typenbildungen liefern. In einem konkreten Test mit fünf Evaluationsinterviews lieferte das System fünf Studierenden-Typen – bei nur fünf Interviews natürlich überdifferenziert, aber inhaltlich nachvollziehbar.

Das Ergebnis fühlt sich beunruhigend nah an einem fertigen Ergebnisbericht an. Die eigentliche Aufgabe der Forschenden wird dadurch anspruchsvoller und nicht wirklich einfacher: Man muss den Datensatz sehr gut kennen, um überhaupt in der Kompetenzrolle zu sein, die Ergebnisse angemessen zu überprüfen. Und die Brille, durch die das Material betrachtet wird, ist vorgegeben durch unsichtbare Prompts – eine spezifische Perspektive auf qualitative Forschung, die andere legitime Brillen vielleicht nicht ermöglicht.

Weitere Tools: KarlAI (vormals DocuMetAI; Schäffer/Lieder, seit 19. Januar 2026 umbenannt) bietet vorgefertigte Prompts für die dokumentarische Methode und weitere, sequenzanalytische Verfahren. Qinsights (Susan Friese) adressiert ihren Verfahrensvorschlag (Conversational Analyse with AI). GoThesis verdient besondere (ich meine das nicht positiv) Erwähnung als Full-Service-Lösung, die sich mit dem Slogan „Abschlussarbeiten fast auf Knopfdruck“ an Studierende richtet: 70.000 monatliche Nutzende, 29,90 €, inklusive einem „Schredder“, der KI-generierte Texte so verändert, dass sie von Detektionstools nicht mehr erkannt werden. Und ein Werbebanner „100 % sicher und legal“ – das erscheint mir spätestens seit Dezember 2025 aufgrund des Gerichtsurteil des Verwaltungsgerichts Hamburg, dass KI-Einsatz bei Leistungsnachweisen auch ohne explizites Verbot als Täuschungsversuch gewertet werden kann, doch etwas, nunja, unangemessen.

### **DSGVO und Anonymisierung: Unhintergehbare Grenzen**

Bei allen genannten Tools ist vor der Nutzung die DSGVO-Konformität zu klären. Die Lage ist komplex: Amazon-Bedrock-Server sind zertifiziert. Das aktuelle Privacy-Framework (dritte Iteration nach zwei erfolgreichen Klagen des österreichischen Anwalts Schrems) steht unter Vorbehalt. Für jede Datenübertragung personenbezogener Daten ist eine Folgenabschätzung erforderlich, die auch von Ethikkommissionen genehmigt werden muss – im Zweifelsfall vermutlich eher zugunsten der Probandinnen.

Die häufig empfohlene Alternative der Anonymisierung ist nach Erkenntnissen des von uns verfassten und in Kürze erscheinenden Artikels (Preprint ca. Februar 2026) keine tragfähige Lösung: Echte Anonymisierung, die Re-Identifikation zuverlässig verhindert, zerstört qualitative Daten bis zur Unbrauchbarkeit. Bereits drei bis vier Merkmalskombinationen ermöglichen KI eine Re-Identifikation von ca. 90 % der Probanden. Das bloße Ändern von Namen und Orten reicht bei weitem nicht aus. Die Empfehlungen von Softwareherstellern, Material „einfach zu anonymisieren“, sind Augenschwermerei. Gleichwohl: In der Praxis wird dies getan! Auffällig ist, dass bei Datenschutzbeauftragten an wenigstens einer bundesdeutschen Hochschulen in den vergangenen Jahren (!) bisweilen nicht eine einzige Nachfrage von Promovierenden zu diesem Thema eingegangen ist.

## **3. Widerstände, Kontroversen und offene Probleme**

Das Thema KI und qualitative Forschung polarisiert und emotionalisiert – in Mailinglisten, auf Tagungen, in Forschungsgruppen. Die Probleme und Reibungspunkte lassen sich in drei Ebenen systematisieren:

**Epistemisch-methodische Ebene:** Bias aufgrund von Trainingsdaten, Halluzination, Varianz des Outputs (nicht reproduzierbar), fehlende intersubjektive Nachvollziehbarkeit, Blackbox-Problem (wie kommt das LLM zu seiner Antwort?), Kontextlimitierung (nicht alle Modelle können 50 Interviews verarbeiten), und eine noch

völlig ungeklärte (oder zumindest nicht breit akzeptierte) methodisierte Integration in bestehende Verfahren qualitativer Textanalyse.

**Akteursbezogene Ebene:** Überforderung durch die Vielzahl an Optionen und Tools, Komplexität des Prompt Engineerings, unklare Agency (bin ich noch die forschende Person?), Risiko unkritischer Übernahme von Inhalten, potenzieller Kompetenzverlust und geringe Lerneffekte.

**Institutionell-rechtliche Ebene:** DSGVO-Problematik, Prüfungsproblematik (Nachweisbarkeit eigenständiger Leistung), unklare Zitierbarkeit, fehlende Standards, ökonomische Fragen (Kosten für Hardware und Lizenzen), ökologische Fragen (Rechenzentren von vielem Quadratkilometern Größe, Atomkraftwerke, die wieder ans Netz gehen), ethische Fragen und rechtliche Unsicherheiten.

Ein konkretes Beispiel für die Polarisierung: Über 400 qualitativ Forschende – vor allem aus dem englischsprachigen Raum – haben in einer Unterschriftenaktion die kategorische Ablehnung von LLMs für rekonstruktive Forschung gefordert. Gegenpositionen kommen unter anderem von Morgan und Friese. Wir sind mitten in einem Aushandlungsprozess.

Auf der anderen Seite der Skala stehen Humanizer-Tools wie Bypass GPT, Higgs Bypass oder Upasai, die KI-generierte Texte so verändern, dass sie von Detektionssoftware nicht mehr erkannt werden. Und es gibt die paradoxe Gegenreaktion: Ein Doktorand in den USA, der redlich arbeitete, aber fälschlich der KI-Nutzung beschuldigt wurde, verschlechtert nun absichtlich seine Texte und zeichnet jeden Tastendruck auf, um seine Autorenschaft belegen zu können (NBC, 28. Januar 2026).

## 4. Kann es trotzdem funktionieren? Einsichten aus einem Re-Analyse-Experiment

Für mich stellte sich immer wieder die empirische Frage: Wäre es überhaupt möglich, wenn jemand die vielen Tipps und Strategien umfassend berücksichtigt, dann eine vollständige qualitative Analyse mit LLMs durchzuführen oder wenigstens substanziell zu unterstützen? Die bisherige Literatur lieferte meist exemplarische Tests mit veralteten Modellen. Ein systematischer Durchstich fehlte und so hab ich mich an einen gewagt: der Re-Analyse einer vor rund 20 Jahren durchgeführten Studie (Qualitative Evaluation). Die Publikation dazu ist in Vorbereitung.

### **Das Ausgangsmaterial**

Als Grundlage diente eine 2007 im VS-Verlag publizierte Studie zur qualitativen Evaluation einer Statistik-Lehrveranstaltung an der Universität Marburg. Zehn Kurzinterviews (je 7–10 Minuten, insgesamt ca. 35–40 Seiten Transkript) wurden von der Fragestellung über Leitfadententwicklung, Interviewdurchführung, Transkription,

fallbasierte Auswertung, deduktiv und induktiver Themenidentifikation bis zum deskriptiven Ergebnisbericht mit Handlungsempfehlungen verarbeitet. Der dokumentierte Gesamtaufwand: 100 Arbeitsstunden. Analytisch lag das Verfahren ähnlich einer qualitativen Inhaltsanalyse – deduktive entlang der Interviewfragen, ergänzt um induktive Befunde, Fallzusammenfassungen und Handlungsempfehlungen für die Dozierenden.

### **Zwei Stränge: Lokal vs. Online**

Lokaler Strang:

Auf einem MacBook Pro mit M3-Prozessor und 128 GB Arbeitsspeicher (der zentrale Engpass für lokales Arbeiten – 48 GB Minimum, besser 96+) wurden Hunderte Prompt-Iterationen durchgeführt: Reverse Prompting, Beispielanker, Parametervariationen, vorheriges Einfügen von Absatznummern ins Material, Decomposed Prompting (Aufgabenzerlegung in separate Threads), nachgeschaltete Prüf-LLMs mit Verifikations-Prompts, drei- bis fünffach iterative Wechselverfahren mit Kreuzvalidierung und vieles mehr. Das Ergebnis: Inhaltlich sind die Auswertungen gut – vergleichbar mit dem Buchkapitel. Aber die Zitatangaben und Zitatdarstellungen sind oft falsch. Lokale Modelle verrutschen in den Absatznummern, liefern falsche Referenzen. Da jedes Zitat manuell überprüft werden müsste – eine extrem lästige Arbeit, die niemand machen wird – ist der lokale Weg für eine vollständige, korrekt referenzierte Analyse derzeit nicht gangbar.

Online-Strang:

Da echte Interviewdaten aus DSGVO-Gründen nicht in Cloud-Dienste hochgeladen werden dürfen, wurden zunächst synthetisierte Interviewdaten erstellt: Aus den Originalinterviews wurden nur die Fragen extrahiert, und mit drei verschiedenen LLMs wurden zehn fiktive Interviews generiert, die in Stil und Umfang den Originaldaten entsprechen (stichprobenartig geprüft). Mit diesen synthetisierten Daten wurde das iterative Prüfverfahren online getestet. Das Ergebnis: Nach einem iterativen Prozess – ein LLM erstellt die Analyse, ein zweites prüft, das Prüfergebnis geht zurück, Überarbeitung, erneute Prüfung, bis nach drei bis vier Durchgängen nichts mehr zu beanstanden ist – waren die Zitate für dieses Datenmaterial zu 100 % korrekt. Richtige Fallnummern, richtige Absatznummern, inhaltlich on point. Das ging aber nur mit den aktuell besten Modellen (Claude Opus 4.5, Gemini 3 pro, GPT 5.2 thinking)

Von der Analyse zur Software

Ein besonderer Aspekt des Experiments: Der gesamte iterative Workflow wurde von mir fortlaufend in einer Word-Datei dokumentiert – welche Prompts, in welcher Reihenfolge, mit welchem LLM, welche Iterationen. Da die von mir verwendeten Prüfprompts generisch waren „bspw. Prüf das bitte nach xyz“ und nicht jeweils individuell, schienen sie mir eine gute Basis zur Automatisierung und Übertragbarkeit auf andere Daten. Das

Worddokument habe ich dann Anthropic Claude Co work übergeben mit der Anweisung: „Mach mir das als Software.“ Und ja, es waren schon ein paar Iterationsschritte, aber nun ist das fertig und voll automatisiert nutzbar.

Das Ergebnis ist eine funktionierende Anwendung mit fünf Schritten: Daten einfügen, Absatznummern automatisch hinzufügen, Fallzusammenfassungen erstellen, deskriptive Analyse durchführen, Synthese und Empfehlungen generieren (jeweils mit 2-5 iterativer Prüfungen), abschließend noch ein finaler Prüfbericht. Bei jeder Absatzreferenz kann per Klick die Originalstelle eingesehen werden. Die Software läuft übrigens gerade während des Vortrags im Hintergrund an einer Auswertung. Die Prompts sind offen und können verändert werden. Kostenpunkt bei 40 Seiten Material rund 5-6 Euro je komplettem Durchlauf.

### Zwischenfazit

Für deskriptive qualitative Evaluation mit klarer Aufgabenstellung ist die vollständige Automatisierung scheinbar technisch im Bereich des möglichen – ohne die natürlich notwendige Einordnung und Abwägung von Maßnahmen und Konsequenzen. Die investierten Arbeitsstunden in Prompt-Entwicklung und Validierung stehen als einmaliger Aufwand dem Ergebnis gegenüber, dass künftige Evaluationen desselben Typs in Minuten statt Wochen durchgeführt werden können. Aber: Die Agency geht verloren, eine eigenständige wissenschaftliche Leistung liegt nicht mehr vor.

Und – entscheidend – dies ist nicht der von mir präferierte Weg. Denn dieser Weg führt dazu, dass Forschende obsolet werden. Mein Herz schlägt aber sowohl für das intensive ausloten, als auch für einen anderen Pfad: selbst forschen, methodisch innovativ, mit KI als Werkzeug statt als Ersatz, wie wir später beim hybriden Interpretieren sehen werden.

## 5. Sechs Studien: Was KI-Nutzung mit uns macht

Was passiert kognitiv, wenn forschende Menschen LLMs verwenden und nutzen sie die Tools überhaupt? Sechs aktuelle Studien – zumeist aus 2025, da ältere KI-Studien sehr schnell zu historischen, nicht mehr aktuellen Dokumenten werden – zeichnen ein differenziertes Bild.

### Studie 1: Flächendeckende Nutzung

Die TU Darmstadt befragte 2025 knapp 5.000 Studierende: 92,6 % nutzen ChatGPT. Nicht alle für qualitative Forschung, aber die Durchdringung ist flächendeckend (vgl. von Garrel & Mayer, 2025: Studie TU Darmstadt).

### Studie 2: Unkritische Übernahme bei fehlender Expertise

Eine Microsoft-finanzierte Studie untersuchte 319 Knowledge Worker an Hochschulen. Zentrales Ergebnis: Menschen ohne Expertise zu einem Thema übernehmen KI-Outputs

überproportional häufig unhinterfragt. Je elaborierter das eigene Domänenwissen, desto eher wird hinterfragt und widerstanden (vgl. Lee et al., 2025: Microsoft Research Paper).

#### Studie 3: Kein Lerneffekt, keine Erinnerung

Drei Studierendengruppen schrieben Essays: mit Papier und Stift, mit Google-Suche (vor KI-Outputs) und mit ChatGPT. Die KI-Gruppe konnte sich 30 Minuten nach Fertigstellung nicht an die Inhalte erinnern. Im MRT zeigte sich: Die Hirnaktivität in lernrelevanten Arealen war bei der KI-Gruppe minimal, bei der Papier-Stift-Gruppe am höchsten. Der LLM-Einsatz ist schneller, aber die Kosten sind ein Bypass des Lernprozesses – keine Verbindung zwischen Produkt und Mensch, kein Kompetenzgewinn (vgl. Kosmyrna et al., 2025: Studie Gehirnaktivität).

Kombination der Studien 1–3: Alle nutzen KI, niemand erinnert sich an die Inhalte, und weil Expertise fehlt, werden KI-Outputs unhinterfragt übernommen. Diese Zuspitzung ist gewiss übertrieben – aber die Richtung stimmt bedenklich.

#### Studie 4: Erosion der Deutungskompetenz auch bei Experten

Hochspezialisierte ÄrztInnen mit etwa zehn Jahren Expertise in der CT-basierten Krebsdiagnostik erhielten KI-Assistenz. Ergebnis: 20 % bessere Erkennungsrate – mehr gerettete Leben. Nach drei Monaten wurde die KI wieder entzogen. Die Erwartung: Rückkehr auf das Ausgangsniveau. Die Realität: Alle Expertinnen fielen um weitere 20 % unter ihr ursprüngliches Niveau. Drei Monate KI-gestützte Arbeit reichten aus, um Deutungskompetenz zu erodieren, die über Jahrzehnte aufgebaut worden war. Das Prinzip „Use it or lose it“ gilt offenbar auch für kognitive Expertise (vgl. Budzyn et al., 2025).

#### Studie 5: Ohne Anleitung kein Mehrwert

Eine weitere Studie aus dem medizinischen Kontext zeigte: KI-Tools erbrachten nur dann einen signifikanten diagnostischen Mehrwert, wenn ihre Einführung didaktisch begleitet wurde. Ohne strukturierte Anleitung blieb der Gewinn aus. (Goh et al., 2024, <https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2825395>)

#### Studie 6: Strukturierter Einsatz mit Reflexionspausen

Die neueste Studie (November/Dezember 2025) belegt: Wenn der LLM-Einsatz strukturiert erfolgt und mit Reflexionspausen sowie aktiver Selbsteinbindung verbunden wird, ergeben sich messbar bessere Ergebnisse als bei unstrukturierter Nutzung. Wer reflektiert und selbst aktiv wird, profitiert. (Gerlich 2025, <https://www.mdpi.com/2306-5729/10/11/172>)

#### Schlussfolgerung aus den Studien

Wer ohne Expertise KI nutzt, gewinnt keine Kompetenz und hat keine Verbindung zu den Ergebnissen. Wer mit Expertise KI nutzt, ist möglicherweise einer Erosion der Deutungskompetenz ausgesetzt. Aber: Didaktische Begleitung und strukturierter,

reflexiver Einsatz könnten als Stellschrauben dienen. Die Konsequenz ist zweifach: Wir brauchen didaktische Integration – nicht nur für Studierende, sondern auch für künftige Lehrpersonen. Und wir brauchen methodische Innovation, die Agency stärkt statt sie aufzugeben. Dieser zweite Pfad ist Gegenstand des zweiten Vortragsteils.

## 6. Methodisierte Integration von LLMs in Varianten qualitativer Forschungsmethodik

Verlassen wir nun die Ebene der Technikbetrachtung – der Prompts, der Tools, der Schwierigkeiten – und wenden uns einer grundlegenden Frage zu: Lässt sich das überhaupt lösen, oder ist es strukturimmanent und unlösbar? Es könnte ja sein, dass wir feststellen müssen: Large Language Models sind zwar beeindruckend, aber es gibt schlicht keine methodisch tragfähige Lösung, um sie in qualitativer Forschung einzusetzen.

Diejenigen, die sich mit dieser Frage wissenschaftlich auseinandergesetzt haben, sind bislang sehr wenige. Im wissenschaftlichen Kontext sind Publikationen das Mittel der Wahl, um zur Diskussion beizutragen. Und wenn wir uns diejenigen Publikationen anschauen, die versuchen, Ideen zu formulieren, wie wir mit den vorhandenen Herausforderungen umgehen können, dann finden wir vielleicht zwei Handvoll. Mehr gibt es im Moment noch nicht an publizierten Vorschlägen, die das nutzen, was wir in qualitativer Forschung bereits kennen und etabliert haben.

Der Dreh- und Angelpunkt sind die qualitativen Forschungsmethoden selbst – ein wunderbar bunter Blumenstrauß an ganz unterschiedlichen, vielfach erprobten und weitgehend akzeptierten Strategien und Grundhaltungen: objektive Hermeneutik, Biographieanalyse, rekonstruktive Analyse qualitativer Daten, Grounded Theory, Metaphernanalyse, Diskursanalyse, qualitative Inhaltsanalyse in ihren verschiedenen Ausgestaltungen und viele weitere. Wir haben vermutlich 40 bis 60 verschiedene kommunizierte und praktizierte Methoden. Und nun stellt sich die Frage: Wie kann diese neue Technik Einzug finden in das, was für uns als qualitativ Forschende besonders wichtig ist? Denn die Methoden sind keine Gängelung, sondern sie sollen die Wissenschaftlichkeit sicherstellen – nicht nur als Schritt-für-Schritt-Anleitung, sondern als Leitlinien dafür, wie Forschungsprozesse zu gestalten und Forschungsfragen mit gegebenem Material zu beantworten sind.

Im Folgenden stelle ich sechs Ansätze vor, bei denen Forschende, die in qualitativer Forschung bewandert sind, darüber nachdenken, wie diese neue Technik mit etablierten Methoden verbunden werden kann. Der älteste stammt von Juni 2024 – fast noch druckfrisch, aber trotzdem bereits der älteste.

Das Wertvolle an all diesen Vorschlägen: Wenn Menschen Vorschläge machen, kann man Dinge ausprobieren, und das Feld kann herausfinden, was vielleicht dauerhaft brauchbar ist – oder vielleicht auch gar nichts davon.

### **1. Dokumentarische Methode mit KI (Schäffer/Lieder, Juni 2024)**

Burkhard Schäffer und Fabio Lieder von der Bundeswehr-Universität München haben eine Publikation vorgelegt, die zeigt, wie die dokumentarische Methode mit KI umgesetzt werden könnte. Die dokumentarische Methode gehört zu den populären Verfahren zumindest im deutschsprachigen Raum, mit Ralf Bohnsack als zentraler Figur. Wer das live erleben möchte, hat dazu Gelegenheit bei vielen Interpretationswerkstätten oder dem Magdeburger Methodenworkshop – der zweitgrößten Veranstaltung im deutschsprachigen Raum mit regelmäßig 17 bis 18 parallelen Workshops und etwa 350 Promovierenden.

Was war an dieser Publikation für mich augenöffnend? Zwei Dinge. Erstens: Ein Prompt muss viel komplexer sein als ein oder zwei Zeilen. Schäffer und Lieder transportieren in ihrer Publikation einen Basis-Prompt von zwei DIN-A4-Seiten Länge. Das war mein erstes Verständnis: Prompts müssen umfangreich sein, und wenn sie umfangreich sind, steuern sie das Large Language Model zielgerichteter an.

Zweitens, und das finde ich mindestens ebenso wichtig: Es gibt bestimmte Strukturelemente in solchen Prompts, die spezifisch für qualitative Forschung vorteilhaft zu sein scheinen. Schäffer und Lieder identifizieren fünf unterschiedliche Elemente: einen Research-Project-Prompt (Informationen über Art und Umfang der Datenerhebung, die zentrale Forschungsfrage, die Zielsetzung der Studie, die Beschreibung der konkreten Untersuchungsgruppe), einen Basic-Theory-Prompt (grundlegende epistemologische Sichtweisen: Wird Wahrheit als objektiv gegeben betrachtet oder konstruktivistisch verstanden?), einen Object-Theory-Prompt (zentrale Konzepte aus dem konkreten Forschungsfeld, Informationen über spezifische Forschungssituationen), einen Methodology-Prompt (die Grundprinzipien der gewählten Methode) und erst dann den Method-Prompt – die eigentliche Schritt-für-Schritt-Anleitung zum inhaltlichen Vorgehen.

Das Entscheidende an dieser Prompt-Architektur ist, dass sie eine Metastruktur darstellt: Wer nicht nach der dokumentarischen Methode arbeitet, sondern etwa rekonstruktive Analyse qualitativer Daten betreibt, kann den Methodology-Prompt entsprechend austauschen. Wenn man mit einem solchen umfassenden Prompt arbeitet und dann konkretes Datenmaterial eingibt – nicht 100 Interviews, sondern Auszüge von einer bis maximal drei DIN-A4-Seiten –, dann ist der Output substanziell näher an etwas, das wir für den Gegenstand angemessen halten, als wenn wir schlicht schreiben: „Mach bitte dokumentarische Methode an diesem Textstück.“

Schäffer und Lieder haben auf Basis dieser Arbeit ihre Software KarlAI entwickelt (vormals DokuMetAI), bei der man über eine Weboberfläche verschiedene methodische

Prompts nutzen kann. Auch dort ist die Frage der Transparenz relevant: Bei DokuMetAI waren die Prompts nicht einsehbar. Ob KarlAI das geändert hat, muss geprüft werden. Wichtig ist auch hier: Die Nutzung ist mit Datenübertragung an Dritte verbunden und erfordert daher eine entsprechende Einwilligungserklärung der befragten Personen.

## **2. Qualitative Inhaltsanalyse mit LLMs (Kuckartz/Rädiker, 6. Auflage)**

Udo Kuckartz und Stefan Rädiker haben in der sechsten Auflage ihres Buches zur qualitativen Inhaltsanalyse ein neues, zehntes Kapitel eingefügt, das auch kostenfrei als PDF herunterladbar ist (mindestens in englischer Sprache). Dort schlagen sie vor, wie die qualitative Inhaltsanalyse mit Large Language Models umgesetzt werden könnte.

Das Besondere an ihrem Ansatz: Sie schärfen das Verständnis dafür, dass es nicht vorteilhaft ist, dem Large Language Model eine zu große Aufgabe zu geben. „Mach die Inhaltsanalyse“ ist so vielgestaltig – mit ganz unterschiedlichen Arbeitsschritten, Perspektiven und Zwischenstopps –, dass man damit das Large Language Model verständlicherweise überfordert. Und in der Analogie: Man würde sehr wahrscheinlich auch eine assistierende Person überfordern, wenn man ihr sagt „Mach mir mal eine Inhaltsanalyse“, ohne weitere Spezifizierung. Die Person würde sinnvollerweise Rückfragen stellen.

Die Kernbotschaft von Kuckartz und Rädiker lautet daher: Eine komplexe Aufgabe muss in unterschiedliche, handelbare Arbeitsschritte zerlegt werden. Für jeden dieser Schritte geben sie einen Prompt an. Insgesamt identifizieren sie 13 Schritte in fünf übergeordneten Phasen: Datenexploration, Kategorienentwicklung, Codieren, Analyse der codierten Daten und Ergebnisbericht schreiben.

Zur Datenexploration gehört etwa die Möglichkeit, sich Zusammenfassungen von Interviews erstellen zu lassen oder Fragen an das Material zu stellen: „Hat Person XY gesagt, dass sie Interesse an einer weiteren Fortbildung hat?“ Das ist Datenerkundung, die das Large Language Model übernehmen kann, ohne dass man alles selbst lesen müsste.

Besonders interessant ist der Codierschritt: Man kann deduktive Codes manuell entwickeln und sie gut operationalisieren – mit Ankerbeispielen, Abgrenzungen zu anderen Themen und Begründungen. Wenn diese Operationalisierung vernünftig gemacht wurde, kann potenziell auch eine andere Person – oder eben ein LLM – das Textmaterial durchgehen und die entsprechenden Stellen identifizieren. In MAXQDA gibt es genau diese Funktion: Wenn die Codes vorhanden sind und die Operationalisierung in den Memos steht, kann das gesamte Material automatisch nach diesen vorhandenen Codes kodiert werden. Für jede Kodierung wird das Modell im Hintergrund gezwungen, eine Begründung zu schreiben, warum es glaubt, dass die Stelle in diesen Code passt. Diese Begründung wird im Kommentar abgelegt.

Das klingt zunächst gut, birgt aber ein ernstzunehmendes Problem. Ein Kollege, der an einer amerikanischen Universität das Qualitative Research Institute leitet, berichtet: Niemand codiert dort mehr selbst. Der Mittelbau überprüft nur noch die automatischen Codierungen – er ist zum Clickworker degradiert worden. Es wird zwar erwartet, dass jede Kodierung überprüft wird, aber das mutet doch fast schon dystopisch. Und dieses Phänomen stellt er nicht nur an seiner Hochschule fest: In den USA fehlen die Begrenzungen der DSGVO, die Leute benutzen die Tools einfach, der Publikationsdruck ist enorm – nicht Jahre, sondern Wochen stehen für eine Studie zur Verfügung.

Auch die Auswirkungen auf die Forschungsförderung sind bereits spürbar: Wo die DFG früher 10.000 Euro für manuelle Transkription ohne Weiteres bewilligte, gibt es heute nur noch einen Bruchteil der Mittel – das geht doch automatisch. Wann kommt die analoge Kürzung für die automatisierte Analyse?

### **3. Query-based Analysis (David Morgan, Januar 2025)**

David Morgan, ein Urgestein der amerikanischen qualitativen Forschung (längst emeritiert), hat im Januar 2025 seinen Ansatz der Query-based Analysis publiziert. Pointiert gesprochen: Codes sind tot. Niemand soll mehr kodieren. Warum haben wir Codes und codieren? Weil unser Gehirn die Datenmenge sonst nicht bewältigt – Codes helfen, diese riesige Datenmenge handhabbar zu strukturieren. Aber genau das könnten doch LLMs übernehmen.

Morgans Grundidee: Wir wollen eine Forschungsfrage beantworten, und wir haben das Material. Dann nutzen wir die LLMs genau dafür – als unser „Brain“. Wir stellen Fragen an das Material, und die LLMs beantworten sie. Kein Codieren mehr.

Der Ansatz hat drei Schritte, immer im Wechselspiel zwischen Mensch und KI: Der Mensch etabliert den Kontext und formuliert breite initiale Fragen – das ist im Grunde das Prompt-Engineering an dieser Stelle. Die KI führt eine erste Textverarbeitung durch, identifiziert möglicherweise Muster und schlägt Zusammenfassungen von Hauptthemen vor. Der Mensch bewertet die KI-Antworten kritisch – was nur gelingt, wenn man seine Daten kennt. Es folgen spezifischere Folgefragen und dann Untersuchungen zur Stützung der gefundenen Muster im Material.

Für mich stellt sich dabei die Frage: Ist es dann wirklich noch ein Zeitvorteil? Und ist Zeitvorteil vielleicht eine ganz falsche Dimension zu fragen? Am Ende steht die Generierung möglichst gut validierter – nicht mathematisch validierter, sondern durch Beispiele validierter – Aussagen zum Material. Man könnte die Query-based Analysis auch als etwas Neues sehen, nicht mehr als Integration in bestehende qualitative Verfahren, sondern als eine sich entwickelnde Form neuer Methodik.

### **4. Hybrides Interpretieren (Krähnke/Pehl/Dresing, Januar 2025)**

Der vierte Ansatz stammt aus unserer eigenen Arbeit. Uwe Krähnke, Thorsten Pehl und ich haben das hybride Interpretieren im Januar 2025 publiziert. Um die Idee zu erklären,

muss ich zunächst noch einmal die Probleme thematisieren, die wir im Umgang mit LLMs sehen. Die Probleme: der inhärente Bias der LLMs (welche Personengruppen sind über- oder unterrepräsentiert in den Trainingsdaten?), die Variationsbreite der Antworten (dieselbe Frage heute ergibt eine andere Antwort als morgen), eine wenigstens unklare Agency-Verteilung (wer analysiert hier eigentlich – ich oder das LLM?), die mangelnde Nachvollziehbarkeit (wir bekommen einfach einen Output – warum genau dieser?), das komplexe Prompt-Engineering und die bis dato fehlende Methodenintegration.

Und jetzt der entscheidende Gedanke: All diese Punkte scheinen auch häufig in Interpretationswerkstätten – und dort sogar gewollt und positiv genutzt zu sein. In einer Interpretationswerkstatt treffen sich zahlreiche Personen, um Material gemeinsam zu analysieren. Jede Person hat einen anderen Bias, eine andere Sozialisation – und genau das wird als Stärke gesehen, weil es eine Bereicherung darstellt, etwa wenn Teilnehmende aus unterschiedlichen Disziplinen kommen. Variationsbreite der Antworten? Natürlich gibt es die auch in einer Werkstatt – ich würde vermutlich unterschiedlich antworten, ob ich den Kaffee schon getrunken habe oder nicht. Aber was die Teilnehmenden tun: Sie erzeugen Perspektiven mit textnahen Deutungen, keine Wahrheiten. In qualitativer Forschung gehen wir davon aus, dass es nicht die eine Wahrheit gibt, sondern wir entwickeln argumentativ unterfütterte Perspektiven. Die plausibelste oder gegenstandsangemessenste oder performativste Argumentation gewinnt – solange, bis jemand sich durch andere Argumente dagegen positioniert. Die Agency liegt bei der forschenden Person, die bestimmt, welche Argumente angenommen, hinterfragt oder widerlegt werden. Alles ist perfekt nachvollziehbar – man kann die ganze Werkstatt aufnehmen und transkribieren, und hat dann ein 40-Seiten-Transkript, das minutiös nachzeichnet, wie man von Hütchen auf Stöckchen gekommen ist. Und es braucht keine megakomplizierte Technik: In der Interpretationswerkstatt sprechen wir miteinander – ein sprachbasierter Austausch für die Interpretationsentwicklung. Schließlich: Werkstätten sind gängige Praxis in interpretativer Forschung, also keine fehlende Methodenintegration.

Unsere Idee ist daher: Wir machen genau das mit Large Language Models. Eine Interpretationswerkstatt, bei der ich als Mensch die Agency-besitzende, einnehmende und ausübende Figur bin. Und ich analysiere nicht zusammen mit einem LLM, sondern mit mindestens dreien – und zwar im (Sprach)stil einer Interpretationswerkstatt.

Konkret: Ich suche mir eine Textpassage aus, gebe eine Ersteinschätzung ab – Informationen, eine erste Deutung, meine Probleme mit der Passage – und gebe dies dem ersten LLM. Das ist so instruiert, dass es sich wie ein Teilnehmer einer Interpretationswerkstatt verhält: textnah, keine Spiegelstriche, nicht gleich theoretisieren. Das erste LLM antwortet. Dann nehme ich mein Material, meine Ursprungsfrage und die erste Antwort und gebe alles dem zweiten LLM – es ist jetzt Teil der Runde und hat mitbekommen, was die anderen bisher gesagt haben. Das zweite LLM wird sich auf das erste beziehen und auf mich. Dann nehme ich beide Antworten,

meine und das Material und gebe alles dem dritten LLM – das wird bereits viel elaborierter sein, weil es die anderen beiden und meine Vorarbeit einbezieht.

Und jetzt bin ich an der Reihe. Das ist nicht fertig, sondern jetzt ist es, als hätte die ganze Runde einmal gesprochen. Jetzt sage ich vielleicht: „Das ist kompletter Schrott" oder „diesen Part finde ich verfolgenswürdig, lass uns dem genauer nachgehen" oder „wie kommst du zu dieser Deutung? Leite das bitte genauer her." Und dann startet die nächste Runde im selben jeweiligen Thread.

Man braucht dafür null Software. Drei Browserfenster mit kostenlosen Accounts bei jedwedem LLM genügen. Man kopiert die Inhalte hin und her (DSGVO beachten!) und dokumentiert alles parallel in einem Word-Dokument. Es gibt auch eine Anleitung dafür und einen Didaktik-Artikel, der demnächst erscheint und zeigt, wie man das in die Lehre einbinden kann. Dann macht man zwei, drei, vier Runden, bis man das Gefühl hat, Sättigung erreicht zu haben oder weil der Account am Ende ist, die Zeit nicht mehr reicht oder weil es sich inhaltlich im Kreis dreht.

Da können schon mal zehn DIN-A4-Seiten Text entstehen. Und wenn jemand sagt „Das ist aber viel Text", dann würde ich dem entgegen: wie ist es denn in einer „echten“ Interpretationswerkstatt? Bei anderthalb Stunden Diskussion hat man mindestens so viel Transkripttext, eher mehr. Es ist keine Beschleunigung, aber möglicherweise eine Demokratisierung oder / und eine Intensivierung der Ausarbeitung, gerade für diejenigen, die interpretative Arbeit noch nicht häufig gemacht haben.

Ein praktisches Beispiel mit einer Interviewpassage verdeutlicht den Prozess. Das erste LLM (Gemini) liefert zunächst eine typische, eher oberflächliche LLM-Antwort über „Dynamik zwischen individueller Handlungsmacht und strukturellen Bedingungen" – eine Antwort, die man sofort in die Tonne kloppen würde. Aber man braucht sich noch nicht damit auseinanderzusetzen: Das zweite LLM (ChatGPT) nimmt jetzt darauf Bezug und beginnt, eine erste neue Facette einzubringen und die Antwort des ersten LLM zu kritisieren. Prima! Und dann kommt Claude und identifiziert einen unklaren Wechsel der Aussage von „wir" zu „sie". Ein Phänomen, dass in acht von zehn Promovierenden-Gruppen, die bisher nicht qualitativ geforscht haben, nicht auf Anhieb gesehen wird (aber durchaus eine sehr entscheidende Stelle im Material darstellt). Das ist für erfahrene rekonstruktive Forschende vielleicht offensichtlich, aber für die vielen, die das noch nicht gemacht haben, eine echte Unterstützung.

Es ist substanziell unterschiedlich, ob man ein LLM oder drei nimmt. Bei dreien machen die sich gegenseitig eine „Blutgrätsche" – man wird herausgefordert, hat unterschiedliche Perspektiven vor sich und muss/kann sich dann positionieren. Rollenspiele mit nur einem LLM (etwa „du bist jetzt Klaus Kinski, du bist Helmut Schmidt, du bist Luhmann") funktionieren weniger gut: Wenn man mit demselben LLM arbeitet, ist das nicht so vorteilhaft. Die englische Version des Papers ist im Druck und enthält ein neues Gütekriterium: die Viabilität. Es geht nicht um Wahrheit, die wir

herausarbeiten, sondern um Gangbarkeit – einen gangbaren Weg zu schlüssigen, nachvollziehbaren Interpretationspositionen – und der, so argumentieren wir, scheint hier vorzuliegen.

### **5. Conversational Analysis with AI (Susan Friese, Mai 2025)**

Susan Frieses Ansatz zeigt eine gewisse Nähe zur Query-based Analysis von Morgan: Auch hier wird in Iterationen und mit Fragen an das Material gearbeitet. Es wird nicht mehr codiert. Die Large Language Models werden als Sparringpartner gesehen – wie eine Kollegin oder ein Kollege, mit der oder dem man zusammen das Material bearbeitet.

Friese identifiziert fünf Arbeitsschritte: Kennenlernen der Daten, vorbereitende Analyse, Fragen stellen, Erkenntnisse synthetisieren und die Analyse auf eine höhere Abstraktionsebene bringen (Konzeptualisierung). Sie hat dazu auch eine Webplattform entwickelt (Qinsights), bei der man sich registrieren und Daten hochladen kann – die dann natürlich auch an Dritte übertragen werden. Der Preprint lag seit Mai 2025 vor; seit kurzem ist auch das finale Paper erschienen, allerdings hinter einer Paywall bei Springer.

### **6. Serendipity-Prompting (Krähnke/Dresing/Pehl, in Vorbereitung)**

Der sechste Ansatz stammt ebenfalls aus unserer Gruppe und unterscheidet sich von allen bisherigen in einem wesentlichen Punkt: In den allermeisten anderen Ansätzen gibt das LLM ein Ergebnis oder Zwischenergebnis aus – ich stelle eine Frage, und dann kommt etwas zurück. Beim Serendipity-Prompting drehen wir das Prinzip um, weil wir glauben, dass es produktiver ist für die menschliche Figur, wenn sie keine Antworten, sondern Fragen gestellt bekommt. Das LLM fragt uns – wir fragen nicht das LLM.

Diese Fragen sollen irritieren, sollen die vielleicht nicht explizierten Grundvorannahmen, die wir mitbringen, aufdröseln, kontrastieren und Impulse geben. Die Idee dahinter ist methodologisch gut begründet: Wenn man das tut, wird Serendipity – also glückliche Zufallsentdeckungen und Abduktion begünstigt. Nicht in jedem Fall garantiert, aber in vielen Fällen gibt es irgendwann Kristallisationsmomente, in denen wirklich Neues entdeckt wird.

Vom pragmatischen Gesichtspunkt her: Stellen wir uns eine Seminarsituation vor, in der Studierende in Kleingruppen eine DIN-A4-Seite Material analysieren. Eine Gruppe stockt – Fragezeichen, gerunzelte Stirn, kein Austausch. Als dozierende Person würde man nicht die fertige Analyse präsentieren, sondern versuchen, über die ersten ein, zwei Sprünge hinwegzuhelfen: „Antwortet die Person eigentlich wirklich auf die Frage? Habt ihr euch das mal genauer angeguckt?“ Das ist wie ein didaktischer Helfer zur Selbsterkenntnis. Und genau so funktioniert Serendipity-Prompting.

## 7. DSGVO: Das Damoklesschwert

Und nun das Thema, das nicht fehlen darf: Wie sieht es mit der DSGVO aus? Hier muss ich ehrlich sein: Meine Brille hat sich in der letzten Woche noch einmal substantiell verändern müssen, und das war unangenehm.

Mein bisheriges Verständnis war, dass DSGVO-Konformität etwas mit der technischen Lösung zu tun hat: Server in Amerika – ausgeschlossen. Nicht von amerikanischen Unternehmen – besser. Server in Deutschland – super. Innerhalb der Hochschule oder zuhause – am allerbesten. Dieses Verständnis war unvollständig.

Die Gesetzesgrundlage ist eindeutig: In dem Moment, wenn Daten übertragen werden – und es spielt keine Rolle wohin, auch an die vertrauenswürdige Uni Göttingen – handelt es sich um eine Übertragung an Dritte. Und wenn eine Übertragung an Dritte stattfindet, braucht es die informierte, schriftliche Einwilligung der befragten Person. Das gilt für Interviewdaten, und es kommt nicht darauf an, ob der Server in Amerika oder in Göttingen steht. Mündliche Einwilligung reicht nicht – das schreibt die Gesetzesvorlage zu 100% vor.

Was heißt „informierte Einwilligung“? Die Person muss darüber informiert werden und es verstehen: „Ich werde die Daten übertragen für die Transkription, ich werde mithilfe des Tools MAXQDA auswerten, ich werde Qinsights verwenden, und die Daten werden an den Server XY von KarlAI übertragen.“ Und das muss vor der Datenerhebung geschehen, weil die Einwilligungserklärung Teil der Erhebungssituation ist. Das Beispiel eines Whistleblowers macht die Dimension greifbar: Wenn ein Dissident mir Informationen gibt und ich sage, die Daten werden nach San Francisco übertragen, hat er zumindest die Chance sich zu überlegen, ob das so eine gute Idee ist. Und spätestens seit Snowden sollten wir uns keine Illusionen machen: Übertragene Daten können und werden gelesen werden.

Der Cloud Act ist ein zusätzliches Unsicherheitsmoment: Die Trump-Administration hat ein neues Gesetz, das erlaubt, Gesetze zurückzunehmen, ohne es kommunizieren zu müssen. Es kann also sein, dass das EU-US Data Privacy Framework bereits aufgehoben ist, aber nur die Amerikaner wissen es (Schremp III).

Bei besonders sensiblen Daten – religiöse Zuordnung, sexuelle Themen, Gesundheitsthemen und weitere – gibt es ein nochmals höheres Schutzniveau. In diesen Fällen gibt es eigentlich keine Alternative zu rein lokaler Verarbeitung.

### **Die großen Modelle**

Bei der Nutzung von ChatGPT, Claude oder Gemini über deren Cloud-Dienste verlassen die Daten möglicherweise den EU-Rechtsraum. Hier gilt: Nur mit entsprechender informierter Einwilligung der befragten Personen, der Rückkopplung mit den Datenschutzbeauftragten und der Ethikkommission in Erwägung ziehen. Ich vermute das aber mindestens einer dieser drei Instanzen ein Verbot erteilt.

## **Hochschulbasierte LLM-Infrastruktur: Die GWDG**

Die Gesellschaft für wissenschaftliche Datenverarbeitung Göttingen (GWDG) betreibt die Academic Cloud, bei der sich jeder – auch ohne Universitätszugehörigkeit – einen kostenlosen Account anlegen kann. Die Daten werden innerhalb der GWDG-Server verarbeitet, keine Informationen werden nach außerhalb übertragen, und seit November ist die Infrastruktur ISO-27001-zertifiziert. Trotzdem: Auch diese Nutzung ist eine Datenübertragung an Dritte und erfordert eine Einwilligungserklärung – wenn auch das Vertrauensniveau ungleich höher ist als bei den großen US-Anbietern. GWDG stellt verschiedene Open-Source-Modelle zur Verfügung – nicht Claude, nicht Gemini, nicht ChatGPT (die sind nicht zum Herunterladen verfügbar), sondern Open-Weight-Modelle mit teilweise kuriosen Namen: Llama 3.1 Sauerkraut LM 70B Instruct, Qwen3 235b oder GPT-OSS 120B. Die Größenangabe (z.B. „120B“) bezeichnet die Anzahl der Parameter – die „PS-Zahl des Autos“. 120 Milliarden Parameter ist im Vergleich zu den proprietären Modellen wenig: Die großen Modelle arbeiten vermutlich mit einer Größenordnung mehr, also im Bereich von 1,2 Billionen Parametern.

Welche Modelle sind für qualitative Textarbeit empfehlenswert? Auf Basis eigener Vergleichstests: Gemma 3 – 27b, Qwen 3 235B, OpenGPT-OS 120B und das neuere GLM 4.7. Diese vier sind für qualitative Textanalyse brauchbar. Das chinesische DeepSeek-Modell sei nicht empfohlen – nicht weil es chinesisch ist, sondern weil es sehr oft und schnell auf die Metaebene springt, was für qualitative Forschung wenig nützlich zu sein scheint.

## **Komplett lokale Verarbeitung**

Wer sagt: „Ich will das komplett offline machen“, braucht einen Rechner mit ausreichend Arbeitsspeicher. Im Moment ist Apple-Hardware in diesem Punkt scheinbar das Ideale, was man kaufen kann – andere Plattformen sind deutlich langsamer bei schlechterem Preis-Leistungs-Verhältnis. Ab 64 Gigabyte RAM aufwärts fängt es an, brauchbar zu werden. Das Modell, mit dem ich lokal arbeite, ist Qwen 3 Next 80B Thinking (und auch das Instruct-Modell). Mit einem Mac Studio M3 Ultra (96 GB) oder einem MacBook Pro M3 Max (128 GB) lässt sich ausloten, was lokal möglich ist. Mehr als 128 GB kann man im Moment nicht kaufen. Da sind aber 4000 Euro schnell weg.

Die Software LM Studio dient als kostenlose Chat-Oberfläche auf dem eigenen Rechner und bietet Zugriff auf Hugging Face – die größte Plattform der Welt für Open-Weight-Modelle mit mittlerweile über einer Million verfügbarer Modelle. Ein Werkstattbericht beschreibt detailliert und sehr „nerdig“, welche Schritte nötig sind, welche Tools kostenlos verfügbar sind und wie man das hybride Interpretieren komplett auf dem Notebook betreiben kann.

Die lokalen Modelle haben durchaus Einschränkungen: Zitate weisen manchmal Fehler auf – keine Halluzinationen im eigentlichen Sinne, aber typische Probleme wie Absatzverschiebungen, Paraphrasen statt exakter Zitate, und bei mehrfach genannten

Sachverhalten wird nur die letzte Nennung zitiert, statt auf alle hinzuweisen. Bei den großen Cloud-Modellen lässt sich das durch zwei-, dreimaliges Hin-und-her-Prüfen auf 100% Genauigkeit bringen. Lokal ist das noch nicht gelöst. Aber die Quintessenz: Alles entwickelt sich weiter. Was jetzt lokal geht, wäre vor zwei Jahren undenkbar gewesen. In einem überschaubaren Zeitraum – ein, drei, fünf Jahre – wird die lokale Kompetenz das Niveau der heutigen Cloud-Modelle erreichen.

Wer eine hochschulweite Infrastruktur aufbauen möchte, kommt allerdings nicht um Server herum. Apple stellt keine Server her, also ist Linux oder Windows mit Nvidia-Hardware nötig – und damit bewegt man sich preislich in einer komplett anderen Dimension. Eine einzelne Nvidia H100 GPU kostet um die 30.000 Euro, und man braucht mehrere davon für ein Cluster, das ein großes Modell betreiben kann. Die Uni Göttingen hat 5 Millionen Euro an Mitteln eingeworben, um eine solche Infrastruktur aufzubauen. Mit einer Million ginge es vielleicht auch, aber nicht mit 5.000 oder 10.000 Euro.

### **Anonymisierung: Kein Ausweg**

Die Strategie „Dann anonymisiere ich eben alles vorher“ klingt naheliegend, funktioniert aber nicht. Das lässt sich anschaulich zeigen: Stellen Sie sich eine Person vor, die die Hände in Form einer Raute hält und einer anderen Person das Vertrauen ausspricht. Das ist anonymisiert – keine Namen, keine Orte – und trotzdem weiß im bundesdeutschen Kontext sofort jeder, dass von Angela Merkel während der Eurokrise die Rede ist. Das ist nur ein kurzer Ausschnitt, der zeigt: Anonymisierung, wenn sie wirklich richtig gemacht wird, also so, wie das Gesetz es verlangt – nämlich so, dass ein Insider (ein Familienmitglied, ein anderer Studierender desselben Seminars, eine andere ProfessorIn desselben Fachbereichs) die Person nicht mehr identifizieren kann – vernichtet so viel Material, dass kaum noch etwas analytisch Brauchbares übrig bleibt.

Ein Schlüsselerlebnis dazu: In einem MAXQDA-Workshop an der FU Berlin war ein Text im Schulungsmaterial – ordentlich pseudonymisiert, andere Namen, andere Orte. Eine Teilnehmerin sagte: „Die kenne ich.“ Unmöglich, die heißt ja anders. „Nein, die hat bei mir gegenüber im Flur gearbeitet. Das ist 20 Jahre her.“ Und als wir weiter scrollten, bestätigten sich ihre Angaben. Scheinbar „harmlose“ Merkmale machen in der Kombination eine Person leicht identifizierbar – das gilt gerade umso mehr für den Einsatz von KI (Palantier lässt grüßen). Selbst die Gerichte sagen: Vollständige Anonymisierung sei nicht möglich.

Unsere Position – dargelegt in einem Artikel, der in den nächsten Tagen erscheint: Weder Pseudonymisierung noch Anonymisierung kann leisten, was die Gesetzesvorlage verlangt. Selbst Inhaltsanalyse kann man nicht mit anonymisierten Interviewtranskripten sinnvoll machen, weil viele der thematischen Informationen, die man für die Analyse braucht, genau die sind, die zur Identifizierung führen können.

## 8. Was kann ich jetzt tun?

Ohne Zweifel: einfach benutzen. Das ist der Dreh- und Angelpunkt. Sonst bleibt es hypothetisch – vom Hörensagen. „Die sollen doch halluzinieren und die sollen dieses und jenes.“ Man muss eigene Erfahrung machen.

Konkret: Accounts anlegen bei der GWDG und eventuell auch bei ChatGPT, Claude, Gemini. Alternativ oder ergänzend: einen Rechner mit viel RAM organisieren, LM Studio herunterladen und experimentieren. Die Nutzung vielfältig testen – spielerisch. Man kann zum Beispiel Helmut Schmidt interviewen: Claude als Helmut Schmidt instruieren, die (zunächst grottenschlechten) Antworten dann einem anderen LLM als „fiesem Theaterkritiker“ geben, und die resultierende Kritik zurück an Claude geben mit der Bitte um Verbesserung. Das Ergebnis kann verblüffend authentisch sein – Flutkatastrophe in Hamburg, Anruf von Henry Kissinger, die Mentholzigaretten sind alle und Loki soll welche holen usw.

Man kann auch Interviewmaterial synthetisch generieren – in Seminaren geht das schnell, und dann hat man Rohdaten, mit denen man alles ausprobieren kann, was man immer schon testen wollte. Man könnte sogar eine Metaanalyse machen, wie LLMs glauben, dass Interviews auszusehen haben.

Vibe Coding ist eine weitere spannende Entwicklung: „Ich kann nicht programmieren, aber ich weiß, was ich will, und lasse den Rest das LLM machen.“ Mit Claude Co work (nicht Claude Code) oder Googles Antigravity kann man als Nicht-Programmierer funktionale Software entwickeln.

### **Kritische Fragen an KI-gestützte Analysetools**

Wer Funktionalitäten wie MAXQDA AI Assist, Tailwind oder andere Dienste nutzen will, sollte mindestens drei Ebenen reflektieren:

Rechtlich-institutionell:

Anonymisierung reicht nicht. Sensible Daten erfordern besondere Sorgfalt. Eine Datenschutz-Folgenabschätzung ist nötig, und mindestens der oder die Datenschutzbeauftragte sollte grünes Licht gegeben haben. Eine bundesdeutsche Universität hatte einmal vier Monate lang alle Forschungsprojekte gestoppt, weil datenschutzrechtliche Fragen ungeklärt waren – 30 Promotionen wurden betroffen.

Methodisch-epistemisch:

Wie gehe ich mit der doppelten Blackbox um? Welches LLM hat gearbeitet, welche Prompts wurden verwendet, welche Modellversion? Wie dokumentiere ich das? Wie folge ich den Zitationsregeln? Und vor allem: Was ist meine methodologische Begründung dafür, KI einzusetzen? Effizienz ist kein methodisches Argument. Und wessen Interpretation(sleistung) ist das dann eigentlich?

Praktisch:

Wie validiere ich 100+ automatisch erstellte Codierungen? Macht das jemand wirklich? Und wenn ja, muss ich das Material dafür sehr gut kennen - Bestätigungsfehler scheinen unvermeidlich. Macht euch klar, es hat aktuell noch Experimentalstatus – wer das nutzt, experimentiert in einem wichtigen Feld, aber es gibt bislang keine systematischen Evaluationsstudien dazu und von flächendeckender Akzeptanz scheinen wir auch weit entfernt zu sein.

### **Transformationsfelder**

Wir stehen vor einer Reihe von Transformationsfeldern allein in der qualitativen Forschung: die Kluft in der KI-Nutzung zwischen Erfahrenen und Neueinsteigern, Veränderungen der Forschungsförderung, die Zukunft des wissenschaftlichen Schreibens, die Transformation wissenschaftlichen Arbeitens insgesamt, veränderte Prozesse allein durch die Transkription, der Umgang mit Qualifikationsarbeiten, der Aufwand des Einarbeitens, die Kompetenzentwicklung, die Integration in die Forschungsmethodik, didaktische Aspekte, Qualitätssicherung, DSGVO, die rasante technologische Entwicklung, infrastrukturelle Voraussetzungen, Kosten und ökologische wie ökonomische Implikationen um nur einige zu nennen.

Und es gibt Auswirkungen darauf, wie wir sprechen und handeln. Eine sprachwissenschaftliche Studie zeigte, dass LLMs den Sprachgebrauch der Menschen verändert haben: Begriffe, die LLMs häufiger verwendet hatten (ohne dass dies in menschlichen Videos auf Youtube sichtbar gewesen wäre), gingen nachweislich in den menschlichen Sprachgebrauch über, wie Videos nach Auftreten allgemein verfügbarer generativer KI nachzuweisen war – und das hat Implikationen nicht nur für Sprache, sondern möglicherweise auch für Perspektiven und Denkweisen.

Es war brutal viel. Und wenn ich ehrlich bin: Es fühlt sich jedes Mal widersprüchlich an. Ich kann zeigen, dass die Automatisierung scheinbar funktioniert – und gleichzeitig bin ich überzeugt, dass genau dieser Weg das kaputt macht, was qualitative Forschung eigentlich ausmacht. Das eigene Ringen mit dem Material, das Deuten, das Sich-Einlassen. Beides ist wahr, gleichzeitig. Und das muss man aushalten.

Was mich trotzdem optimistisch stimmt: Es gibt Wege dazwischen. KI als Gegenüber statt als Ersatz. Strukturen, die Reflexion erzwingen, statt sie zu umgehen. Die Bereitschaft, langsamer zu sein als technisch nötig – weil Verstehen nun mal Zeit braucht. Und die Chance, die Sie als Pädagoginnen und Pädagogen haben: Sie können mitgestalten, wie Lernprozesse mit diesen Werkzeugen didaktisch sinnvoll aussehen – denn genau das zeigen die Studien als eine der entscheidenden Stellschrauben hilfreicher Implementation.

Die Transformation findet statt, ob wir das gut finden oder nicht. Es braucht viele Köpfe, hier mitzugestalten, Wege aufzuzeigen und auszuloten: Machen Sie mit!