

## Verhaltenstherapie

Verhaltenstherapie , DOI: 10.1159/000549546

Received: May 15, 2025

Accepted: November 10, 2025

Published online: November 24, 2025

### **Zum Diktat, Language Model: Transkription, Sprecher\*Innenzuordnung und Verarbeitung von Therapiegesprächen mit KI**

Taubitz F-S, Sehn LHA, Alpers GW

ISSN: 1016-6262 (Print), eISSN: 1423-0402 (Online)

<https://www.karger.com/VER>

Verhaltenstherapie

#### Disclaimer:

Accepted, unedited article not yet assigned to an issue. The statements, opinions and data contained in this publication are solely those of the individual authors and contributors and not of the publisher and the editor(s). The publisher and the editor(s) disclaim responsibility for any injury to persons or property resulting from any ideas, methods, instructions or products referred to the content.

#### Copyright:

This article is licensed under the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC) (<https://karger.com/Services/OpenAccessLicense>). Usage and distribution for commercial purposes requires written permission.

© 2025 The Author(s). Published by S. Karger AG, Basel

**Titel:**

**Language Model, bitte zum Diktat: Transkription, Sprecher\*innenzuordnung und datenschutzkonforme Verarbeitung von Therapiegesprächen mit KI**

**Language Model, Take Dictation Please: Automatic Transcription, Speaker Diarization, and Processing of Psychotherapy Sessions with AI in accordance with data protection**

**Kurztitel:**

**KI-gestützte Transkription von Therapiegesprächen**

**AI-Based Transcription of Therapy Sessions**

Friedrich-Samuel Taubitz<sup>1,2</sup>, Linus H. A. Sehn<sup>1</sup>, Georg W. Alpers<sup>1,2</sup>

<sup>1</sup>Department of Psychology School of Social Sciences, University of Mannheim

<sup>2</sup>Otto Selz Institute, University of Mannheim

alpers@uni-mannheim.de

**Korrespondierender Autor**

Korrespondenz zu diesem Artikel ist zu richten an Prof. Dr. Georg W. Alpers, Universität Mannheim, Fakultät für Sozialwissenschaften L 13, 17 68161 Mannheim.

E-Mail: alpers@uni-mannheim.de

## **Abstract**

**Hintergrund:** Neueste Entwicklungen Künstlicher Intelligenz (KI), insbesondere die automatische Transkription und Weiterverarbeitung von Sprache durch Large-Language-Models, eröffnen enormes Potenzial für die Psychotherapieforschung und Praxis. Da Sprache das zentrale Medium der Psychotherapie ist, bieten sich KI-gestützte Verfahren zur Transkription, Analyse und Weiterverarbeitung an – insbesondere mit Blick auf die aufwändigen Dokumentationspflichten. Verfügbare Anwendungen sind jedoch oft ungeeignet, weil Datenschutz und Zuverlässigkeit nicht gewährleistet sind.

**Material und Methoden:** Wir beschreiben und erproben lokal und vom Netzwerk isolierte (air-gapped) - also datenschutzkonforme - KI-Anwendungen im psychotherapeutischen Kontext. Dazu werden innovative Methoden eingeführt, um die Transkriptionsgenauigkeit und die Zuordnung von Sprecher\*innen zu optimieren. Die erforderlichen technischen und methodischen Schritte zur lokalen und automatisierten Aufbereitung, Transkription und die Zuordnung von Sprecher\*innen werden dargestellt. Deren Funktion wurde in einer Pilotstudie anhand simulierter Anamnesegespräche von Psychologiestudierenden mit Schauspielpatient\*innen erprobt, und die Leistungsfähigkeit wurde mit etablierten Methoden (konventionelle Software bzw. manuelle Transkription) verglichen. Zuletzt wurden eine kurze Zusammenfassung der Gespräche sowie Diagnosevorschläge durch ein lokales Large Language Model erstellt.

**Ergebnisse:** Insgesamt erzielten die adaptierten KI-gestützten Transkriptionen eine sehr gute Genauigkeit im Vergleich zu manuellen Transkriptionen und übertrafen größtenteils etablierte Software. Die Zuordnung der Sprecher\*innen (Diarization) erfolgte ebenfalls zuverlässig. Die generierten Zusammenfassungen waren kohärent und fachlich relevant und die Diagnosevorschläge nachvollziehbar.

**Diskussionen und Schlussfolgerungen:** Die explorative Arbeit demonstriert, wie psychotherapeutische Gespräche mit alltagstauglicher, lokaler und datenschutzkonformer Technologie automatisch protokolliert und weiterverarbeitet werden können. Dies kann Routineaufgaben erleichtern und neue Forschungsmöglichkeiten eröffnen.

**Keywords:** Datenschutzkonforme KI, Sprachverarbeitung in der Psychotherapie, KI-Transkription Therapiegespräche, Zusammenfassung mit großen Sprachmodellen

## **Abstract**

**Introduction:** Recent developments in artificial intelligence (AI), particularly automatic transcription and processing of speech using large language models, are expanding the potential for psychotherapy research and practice. As language is the central vehicle of psychotherapy, AI-supported tools are ideal for transcription, analysis, and processing, especially with regard to laborious documentation requirements. However, current applications are often unsuitable because of insufficient data protection and reliability.

**Methods:** We introduce and test on-premise AI applications that can be deployed in air-gapped (network-isolated) environments and are therefore compliant with data protection requirements in the context of psychotherapy. Transcription and assignment to speakers is optimized with innovative approaches. We present the necessary steps for local and automated processing, transcription, and speaker diarization. Simulated anamnestic interviews between psychotherapy trainees and actor-patients were recorded, automatically and locally processed, transcribed, and assigned to the speakers. The results were compared with established procedures (conventional software and manual transcription). Finally, a brief summary and suggestions for diagnosis were generated using a local large-language model.

**Results:** Overall, the adapted AI-supported transcriptions were very good compared to manual transcriptions and mostly outperformed established software. Speaker diarization was reliable. The summaries were coherent and meaningful, and the diagnostic suggestions were reasonable.

**Conclusion:** The pilot study demonstrates how dialogue in psychotherapy can be automatically logged and further processed using technology that is run locally and compliant with data protection. This may facilitate routine tasks and open new opportunities for research.

**Keywords:** Artificial Intelligence (AI) and Data Protection Compliance, Language processing in psychotherapy, AI based transcripts of psychotherapy sessions, Large Language Model summaries

## Theoretischer Hintergrund

Sprache ist das zentrale Medium der Psychotherapie und es wurde längst vorgeschlagen, das Gesprochene aufzuzeichnen und systematisch weiterzuverarbeiten [Rogers, 1942; Lasswell, 1935]. Erst die jüngsten Fortschritte in der Entwicklung Künstlicher Intelligenz (KI) eröffnen ganz neue Möglichkeiten, die bestehenden Hindernisse ökonomischer (Weiter-)Verarbeitung von Sprache in der Psychotherapieforschung und -praxis zu überwinden. Die Integration von KI in der Psychotherapie verspricht nicht nur eine Erleichterung administrativer Aufgaben, sondern auch eine Steigerung der Therapieeffizienz und -qualität [Lutz et al., 2023; Lutz et al., 2024]. Insbesondere die Fortschritte im Bereich der Verarbeitung natürlicher Sprache (Natural Language Processing, NLP) ermöglichen Ansätze, die Effizienz und Qualität therapeutischer Prozesse zu steigern, nicht nur auf einer reinen funktionalen unterstützenden Funktion heraus, sondern auch durch wissenschaftliche auswertbare Detaileinblicke in deren Prozesse [Cummins et al., 2019; Eberhardt et al., 2024; Ewbank et al., 2019; Ewbank et al., 2021].

Die Nachfrage nach psychotherapeutischer Unterstützung ist in den letzten Jahren gestiegen, während die Ressourcen begrenzt bleiben, was zu längeren Wartezeiten und einer gesteigerten Belastung für Psychotherapeut\*innen führt [Singer et al., 2022; Strömsdörfer, 2023; Wittberg et al., 2020]. Neben diesen strukturellen Herausforderungen belastet ein wachsender administrativer Aufwand die therapeutische Arbeit, sodass ein nicht zu vernachlässigender Zeitaufwand für die Protokollierung, Berichterstellung und andere administrative Tätigkeiten entfällt, was wertvolle Ressourcen bindet und von Behandler\*innen häufig als Belastung wahrgenommen wird [Henning et al., 2021; Piechotta, 2008]. KI-Systeme bieten in diesem Bereich eine vielversprechende Entlastung. Obwohl diese Technologien derzeit noch eine gründliche menschliche Nachkontrolle erfordern, wecken die Fortschritte berechtigten Optimismus: So ermöglichen moderne Tools die automatische Erstellung von Gesprächsprotokollen und deren sinnvolle Zusammenfassung [Miner et al., 2020; Zhou et al., 2018].

Neben der potenziellen Entlastung und Qualitätsoptimierung im Bereich der Dokumentation eröffnen zuverlässige und detaillierte Transkriptionen von Therapiegesprächen ein erhebliches Potenzial für die Analyse von Psychotherapieprozessmerkmalen, wie zum Beispiel der detaillierten Auswertung verwendeter Wörter [Alpers et al., 2005], Metaphern oder Interaktionsmuster [Buchholz und Kächele, 2017] womit eine sehr alte Idee aufgegriffen wird, dass wörtliche Transkriptionen das Verständnis therapeutischer Mechanismen vorantreiben kann [Lasswell, 1935]. So lassen sich durch linguistische Analysen und die Identifikation semantischer Muster beispielsweise Anzeichen psychischer Veränderungen wie depressives Erleben oder Symptomverschlechterungen erkennen bzw. vorhersagen [Hur et al., 2024]. Darüber hinaus ist es möglich, über die Gesprächsinhalte (und andere sprachliche Merkmale) Informationen über die therapeutische Beziehung abzuleiten, was dabei unterstützen könnte, Therapien anzupassen oder Therapieentscheidungen zu fördern [Lalk et al., 2024; Lin et al., 2022a]. Die Menge an gewonnenen Informationen könnte mittels Methoden des maschinellen Lernens zur Vorhersage des Therapieerfolgs integriert werden und in ihrer Interaktion mit anderen diagnostischen Daten zur Erfolgsvorhersage beitragen [Taubitz et al., 2022]. Solche Systeme bieten außerdem das Potenzial, während der Behandlung Echtzeitempfehlungen zu geben [Lin et al., 2022b; Lutz et al., 2024].

Die gegenwärtigen Entwicklungen automatisierter Sprachverarbeitung sind vielversprechend, befinden sich derzeit jedoch noch im prototypischen Stadium und sind noch nicht einsatzreif für den Alltag. Es bleiben erhebliche Herausforderungen und zentrale (Weiter-)Entwicklungsfelder, insbesondere vor einer praxisnahen Implementierung. Eine zentrale Anforderung an die Integration von KI-Systemen in Wissenschaft und Praxis betrifft die besondere Schutzbedürftigkeit der in der Therapie behandelten Inhalte sowie die damit verbundenen hohen Anforderungen an den Datenschutz [Alpers et al., 2020]. Frühere KI-Anwendungen mussten aufgrund ihrer enormen Rechenleistungsanforderungen hauptsächlich in Rechenzentren von Drittanbietern (z.B. Amazon Web Services™, Google Cloud Platform™, Microsoft Azure™) betrieben werden. Die besonders schutzwürdigen Daten von Psychotherapiepatient\*innen verbieten jedoch deren Speicherung und Verarbeitung auf Servern solcher - oft sogar ausländischer - Drittanbieter [Alpers et al., 2020]. Aktuelle technologische Entwicklungen ermöglichen die lokale Verarbeitung sensibler Daten auf institutionseigener Hardware - auch auf handelsüblichen Systemen - wodurch Risiken wie unbefugter Zugriff durch Dritte, Datenschutzverletzungen und der Verlust der Kontrolle über die Datenverwendung sinnvoll minimiert werden können. Solche lokalen Lösungen können bei besonders hohen Sicherheitsanforderungen auch getrennt vom Netzwerk (air-gapped) eingesetzt werden, um das Risiko unbefugten Zugriffs weiter zu verringern. Diese Fortschritte ermöglichen eine datenschutzkonforme Nutzung von Methoden zur

Transkription und Weiterverarbeitung von Therapiegesprächen unter vollständiger Datenhoheit der Praxis oder Forschungseinrichtung. Die Wahl einer geeigneten technischen Lösung erfordert dabei eine sorgfältige Abwägung institutioneller Anforderungen an Skalierbarkeit (z. B. Anzahl der Nutzer\*innen), Nutzerfreundlichkeit und Sicherheit.

Neben datenschutzrechtlichen Aspekten haben bisherige Anwendungen methodische Einschränkungen. Beispielsweise erreichen KI-Systeme im Bereich der Transkription und der Zuordnung der Sprecher\*innen unter idealen Bedingungen eine hohe Genauigkeit, sind jedoch in komplexeren, praxisnahen Szenarien teilweise weniger zuverlässig [Siddharth und Beigi, 2022].

Darüber hinaus haben gängige Modelle und Methoden (zum Beispiel die Whisper-Modellreihe von OpenAI™) Probleme mit dem Konfabulieren falscher und neuer Informationen [Koenecke et al., 2024]. Der Begriff „Konfabulation“ wird anstelle von „Halluzination“ verwendet, da Sprachmodelle keine bewussten sensorischen Wahrnehmungen haben, sondern – analog zu menschlichen Konfabulationen – fehlerhafte Rekonstruktionen von Informationen auf Basis gelernter Muster generieren, ohne diese als inkorrekt zu erkennen [Smith et al., 2023]. Fehltranskriptionen, wie Konfabulationen oder Auslassungen, können die Validität der Ergebnisse, gerade im medizinischen und therapeutischen Kontext, kritisch beeinträchtigen. Schon kleine Veränderungen oder konfabulierte Informationen, zum Beispiel Aussagen über gewalttätige oder sexuelle Handlungen können die Bedeutung eines Gesprächs erheblich verändern [vgl. Koenecke et al., 2024]. Diese Veränderungen können wiederum eine daraus resultierende Zusammenfassung verfälschen, was Risiken für die Sicherheit und Akzeptanz solcher Technologien mit sich bringt.

Eine weitere Herausforderung betrifft die Generalisierbarkeit der diskutierten Befunde. Die verwendeten Modelle und Methoden sind häufig sprachabhängig, was dazu führt, dass Erkenntnisse zu ihrer Performanz aus einem englischsprachigen Kontext nicht immer auf einen deutschsprachigen Anwendungskontext übertragbar sind.

Zusammengefasst bieten KI-Methoden das Potenzial, sowohl die Wissenschaft als auch die Praxis zu unterstützen und zuvor schwer umsetzbare Möglichkeiten zu eröffnen. Gleichzeitig befinden sich die gegenwärtigen Entwicklungen noch in den Anfängen und bergen konkrete Risiken, sodass eine stetige Auseinandersetzung mit und eine Weiterentwicklung der Methoden erforderlich ist.

Im Rahmen dieser explorativen pilotierenden Machbarkeitsstudie zum Einsatz innovativer, ressourcenschonender und datenschutzkonformer KI-Systeme im psychotherapeutischen Kontext, war es das Ziel, anhand simulierter Erstgespräche praktische Lösungsvorschläge für zentrale Herausforderungen der automatisierten Transkriptgenerierung aufzuzeigen. Dazu gehören insbesondere eine hohe Genauigkeit der Transkription, eine präzise Zuordnung der Sprecher\*innen, sowie die strukturierte Zusammenfassung der Gesprächsinhalte für Protokollzwecke. Zusätzlich soll evaluiert werden, inwieweit aus den generierten Transkripten von lokal betreibbaren Large Language Models diagnostische Empfehlungen abgeleitet werden können.

Neben dieser Machbarkeitsdemonstration zielt die Arbeit auch darauf ab, das Verständnis für solche Methoden zu fördern und einen fundierten Zugang zu den unweigerlich aufkommenden Entwicklungen in diesem Bereich zu erleichtern.

## Methode

### Audio-Material und Vorgehen

Die vorliegende explorative Studie basiert auf vier simulierten psychotherapeutischen Erstgesprächen, welche im Rahmen eines Masterseminars zur Gesprächsführung in der Psychotherapie („Psychotherapiemaster“) zwischen Masterstudierenden und Schauspielpatient\*innen durchgeführt wurden [Alpers und Hengen, 2021; Alpers und Steiger-White, 2020; Fraunfelter et al., 2025]. Alle Teilnehmenden erteilten nach Aufklärung ihre schriftliche, informierte Einwilligung. Das Setup orientierte sich an einem psychotherapeutischen Erstgespräch und ermöglichte die Elaboration unter datenschutzrechtlich unbedenklicheren Bedingungen. Die transkribierten Sitzungen waren durchschnittlich 33 Minuten lang (kürzestes Gespräch = 27 Minuten, längstes Gespräch = 42 Minuten).

In der Studie wurden drei unterschiedliche Dyaden untersucht: eine weibliche Studierende als Therapeutin mit einem männlichen Schauspielpatienten, zwei weibliche Studierende als Therapeutinnen mit jeweils einer weiblichen Schauspielpatientin sowie ein männlicher Studierender als Therapeut mit einer weiblichen Schauspielpatientin.

Zur Untersuchung des Einflusses verschiedener Mikrofon-Setups auf die Transkriptionsqualität wurden unterschiedliche Anordnungen eingesetzt. Als Basis-Mikrofon-Setup (BM) dienten Tisch- bzw. Webcam-

Mikrofone. Weiterhin kam das omnidirektionale Mic 2 der Firma DJI™ (im Folgenden als Premium-Funkmikrofon bezeichnet) zum Einsatz – entweder direkt an der Oberbekleidung befestigt (Lavalier-Setup; PFM-L) oder auf dem Tisch (PFM-T) positioniert. Alle Aufnahmen mit diesem Premium-Funkmikrofon erfolgten im Stereoformat, sodass zwei separate Tonspuren entstanden. Alle Sitzungen wurden demnach einmal mit dem Basis-Mikrofon-Setup (BM) aufgezeichnet, welches eine Mono-Tonspur erzeugte, sowie einmal mit einem Premium-Funkmikrofon-Setup (entweder im Lavalier-Setup oder auf dem Tisch platziert), woraus sich 8 Audioaufnahmen 4 unterschiedlicher Erstgespräche ergaben.

### Aufbereitung & Transkription

Die Erstellung verwertbarer Transkripte therapeutischer Gespräche erfordert mehrere aufeinander aufbauende Schritte: die Audioaufnahme selbst, deren Qualität durch die Wahl des Aufnahmesystems beeinflusst wird, die automatisierte Transkription der gesprochenen Sprache sowie die Zuordnung der Äußerungen zu den jeweiligen Sprecher\*innen der therapeutischen Dyaden. Während für diese Teilaufgaben verschiedene technische Lösungen existieren, die sich in Komplexität und Ressourcenbedarf unterscheiden, fehlen systematische Vergleiche ihrer Leistungsfähigkeit im psychotherapeutischen Kontext, weshalb die vorliegende Untersuchung unterschiedliche Methodenkombinationen evaluiert.

Alle 8 Audioaufnahmen (4 Dyaden, zwei Mikrofon Setups) wurden transkribiert. Bei der Transkription wurde OpenAIs Standard-Implementation, in der Standard-Hyperparametereinstellungen (siehe Appendix A) von *whisper* in Python [Radford et al., 2022] mit dem Transkriptionssystem *verbatim* [Petit und Sehn, 2025] verglichen, welches im Rahmen dieser Arbeit erweitert und für die ressourcenschonende, lokale Nutzung angepasst wurde. Letzteres nutzt zwar ebenfalls die *whisper*-Modelle, nimmt aber eine intelligente Anpassung der zu transkribierenden Zeitfenster vor (s.u.). Nur die 4 Audioaufnahmen des Premium-Funkmikrofons wurden mittels unterschiedlicher Arten den Sprecher\*innen zugewiesen (Deep-Learning-Analyse mit *pyannote* vs. Lautstärkeanalyse der Stereospuren) und transkribiert.

Sämtliche Prozesse und Anwendungen wurden lokal, auf einem handelsüblichen Laptop (2021 MacBook Pro M1 Max 32GB; Apple™) durchgeführt; d.h. es wurden keine Daten an Dritte (wie bei der Verwendung externer Server) weitergegeben oder für weitere Trainingszwecke von Sprachmodellen zur Verfügung gestellt. Die lokale Verarbeitung wurde bewusst gewählt, um zu demonstrieren, dass datenschutzkonforme Transkription und Analyse ohne Weitergabe sensibler Therapiedaten an Dritte möglich sind und vollständige institutionelle Datenhoheit gewährleistet werden kann. Die gewählte handelsübliche Hardware soll exemplarisch aufzeigen, dass entsprechende Lösungen ohne hochspezialisierte Workstations oder Server-Infrastruktur umsetzbar sind. Eine schematische Darstellung der beschriebenen Prozesse findet sich in Abbildung 1. Die Transkription von Therapiegesprächen beginnt mit der Nachbearbeitung der Aufnahmen um die Stimmen bei Bedarf von Hintergrundgeräuschen zu isolieren und den Lautstärkepegel zu normalisieren. Für die Isolation können verschiedene Ansätze genutzt werden [z.B.: Demucs, Défossez et al., 2019 oder FRCRN, Zhao et al., 2024]. Für die Lautstärkenormalisierung wurde sowohl der Standard R128 [European Broadcasting Union, 2023] sowie der im Programm FFmpeg [FFmpeg Developers, 2025] enthaltene Filter *spechnorm* verwendet. Im nächsten Schritt wurde entweder *verbatim* [Petit und Sehn, 2025] oder OpenAIs Standard-Implementation verwendet, um die Transkription und, wenn möglich, die Zuordnung der Sprecher\*innen durchzuführen.

### Abbildung 1.

*Verbatim* arbeitet mit einem dynamisch wachsenden Zeitfenster zur Audioverarbeitung, der anders als herkömmliche Systeme funktioniert, die oft fest definierte Fenster (z.B. 30 Sekunden) verwenden. Zunächst werden Audiodaten in ein anwachsendes Fenster geladen, wobei die Sprachaktivitätserkennung (*Silero VAD*) [Silero Team, 2024] Episoden mit Stille am Anfang des Fensters identifiziert und überspringt. Der methodische Vorteil besteht darin, dass er sich auch für Echtzeitverarbeitung eignet, was zwar im Rahmen dieser Studie nicht genutzt wurde, aber bei der Anwendung von Transkriptionssystemen im Praxisalltag relevant wäre. Ein weiterer Vorteil besteht darin, dass die Zeitfenster den Trainingsdaten aktueller KI-Transkriptionsmodelle nahekommen, die typischerweise 15-30 Sekunden lange Aufnahmen mit maximal 2-3 Äußerungen umfassen und nicht mitten im Satz enden.

Die Transkription erfolgt in mehreren Durchläufen, wobei die sukzessiv wachsenden Audioabschnitte in *verbatim* analysiert werden. Sprachsegmente werden zunächst als *unconfirmed* (unbestätigt) markiert und erst nach mehrfacher konsistenter Erkennung in nachfolgenden Transkriptionen der größer werdenden Fenster zu *confirmed* (bestätigt) hochgestuft. Sobald die zu semantisch sinnvollen Äußerungen (*utterances*) gruppierten Wortketten einer bestätigten Transkription die Minimallänge von einer Sekunde überschritten

haben und mit einem Satzzeichen enden, werden sie als *acknowledged* (anerkannt) markiert und in die endgültige Transkription übernommen. Erst dann verschiebt *verbatim* das Audiofenster zum Ende der letzten anerkannten Äußerung und beginnt den Prozess für den nächsten Abschnitt. Zusätzlich werden von *verbatim* die identifizierten Sprachsegmente den jeweiligen Sprechern zugeordnet. Hierfür stehen zwei Ansätze zur Verfügung, welche je nach Voraussetzung unterschiedliche Vorteile mit sich bringen:

Der erste Ansatz [*pyannote*; Bredin, 2023; *Plaquet* und Bredin, 2023] basiert auf KI-gestützten Technologien, insbesondere auf Deep-Learning-Modellen zur Extraktion charakteristischer Stimmmerkmale („Speaker Embeddings“). Diese Embeddings werden anschließend mithilfe von Clustering-Verfahren, wie dem „Agglomerative Hierarchical Clustering“ ausgewertet, um einzelne Sprecher\*innen voneinander zu unterscheiden. Die Methode liefert besonders zuverlässige Ergebnisse bei deutlich unterschiedlichen Stimmen, da die Sprecher\*innen klar voneinander abgegrenzte „Embeddings“ aufweisen. Allerdings stößt dieser Ansatz an seine Grenzen, wenn sich die Stimmen der Beteiligten stark ähneln, wodurch es zu einer ungenauen oder fehlerhaften Zuordnung kommen kann.

Der zweite Ansatz basiert auf der Energieanalyse von Stereo-Audiospuren zur Sprechererkennung. Statt komplexer Sprachmustererkennung wird die Energieverteilung zwischen den beiden Stereokanälen untersucht, indem für kurze Zeitfenster (standardmäßig 0,5 Sekunden) die kumulierte Signalstärke - also die Summe der Absolutwerte der Amplituden - in jedem Kanal berechnet wird. Durch Vergleich dieser Energiewerte werden Energieverhältnisse zwischen linkem und rechtem Kanal ermittelt und mit anpassbaren Schwellenwerten verglichen, um dominante Sprecher\*innen zu bestimmen. Bei nahezu identischen Energiewerten wird zusätzlich die maximale Amplitude als Entscheidungskriterium herangezogen. Die Methode nutzt somit die natürliche räumliche Trennung der Audiosignale, die durch unterschiedliche Distanzen der Sprecher\*innen zu den jeweiligen Mikrofonen entsteht, und ermöglicht eine effiziente Zuordnung von Sprechpassagen ohne komplexe akustische Modellierung.

Die erstellten Transkripte wurden mit menschlich erstellten Transkriptionen verglichen und die Word Error Rate (WER) für die Transkription und Word Diarization Error Rate (WDER) für die Zuordnung der Sprecher\*innen bestimmt. Zur Berechnung der Metriken wurden das von Wang et al., [2024] vorgeschlagene Framework verwendet.

Die manuellen Transkripte wurden von drei unabhängigen Rater\*innen angefertigt. Jedes Transkript wurde dabei von einer Person erstellt und anschließend im Konsensverfahren abgeglichen. Bei Unstimmigkeiten erfolgte eine gemeinsame Abstimmung, um eine möglichst hohe Konsistenz des Vorgehens über alle Transkripte hinweg sicherzustellen.

Bei der Erstellung der menschlichen Transkripte wurde jedes Wort transkribiert, auch Einschübe (wie zum Beispiel „Ja“, „Achso“). Silben- und Wortwiederholungen („ja, ja, also ja“), wurden erfasst und Wortverschleifungen („so‘ n bisschen“) wurden geglättet (zu „so ein bisschen“) um eine Standardisierung zu ermöglichen. Protosprachliche Lautäußerungen (wie zum Beispiel „mh“) wurden nicht erfasst, da diese bei den angestrebten Zusammenfassungen nicht integriert werden könnten, und davon ausgegangen werden kann, dass die Validität hiervon weitestgehend unbeeinträchtigt bleibt [Naffah et al., 2025]. Die verwendeten Transkriptionsmodelle wurden nicht primär für die Erkennung solcher paralinguistischer Merkmale optimiert, sodass deren Einbeziehung zu einer systematischen Verzerrung der Fehlerraten geführt hätte. Dies bedeutet jedoch nicht, dass solche sprachlichen Phänomene in der psychotherapeutischen Kommunikation irrelevant wären; vielmehr existieren derzeit noch keine ausgereiften technischen Lösungen, um diese systematisch und zuverlässig zu erfassen und für die Analyse nutzbar zu machen. Eine vertiefte Auseinandersetzung mit diesem Aspekt liegt jedoch außerhalb des Fokus der vorliegenden Arbeit.

### Weiterverarbeitung und Aggregation

Nach der Transkription der Audiospur in eine Textdatei erfolgt eine Weiterverarbeitung des generierten Textes mithilfe einer ebenfalls lokalen, und somit netzwerkisoliert betreibbaren Pipeline namens *concisum* [Sehn, 2025], die auf der systematischen Verarbeitung der generierten Transkripte durch ein Large Language Model basiert. *Concisum* nutzt zwei aufeinander aufbauende System-Prompts, um in einem mehrstufigen Verfahren eine Zusammenfassung zu generieren. Beide Prompts dienen als rollen- und stildefinierende Instruktionen innerhalb einer hierarchisch organisierten Zusammenfassungspipeline

Der erste Prompt („chunk summarizer“) weist das Sprachmodell an, einzelne Abschnitte des Transkripts aus der Perspektive eines fachkundigen „Experten für die Zusammenfassung psychotherapeutischer Sitzungen“ zu kondensieren. Dabei sollen zentrale Inhalte, therapeutische Interventionen sowie relevante Gefühle,

Gedanken und Verhaltensmuster prägnant und professionell in der dritten Person zusammengefasst werden.

Der zweite Prompt („full summarizer“) integriert die zuvor erstellten Teilzusammenfassungen in eine kohärente Gesamtzusammenfassung der gesamten Sitzung. Diese erfolgt in Form einer strukturierten Synthese, die den Themenverlauf, therapeutische Techniken sowie Schlüsselmomente der Sitzung sachlich und präzise in maximal 300 Wörtern abbildet.

Parallel zur Zusammenfassungspipeline implementiert *concisum* eine eigenständige Diagnose-Pipeline, die auf zwei spezialisierten Agenten basiert. Der erste Agent („symptom extractor“) fungiert als „Experte für die Identifikation psychologischer Symptome“ und extrahiert systematisch klinisch relevante Anzeichen aus den einzelnen Transkript-Segmenten. Dabei werden Stimmungsprobleme, kognitive Veränderungen, Verhaltensmuster, physiologische Symptome und soziale Beeinträchtigungen identifiziert. Für jedes erkannte Symptom werden Name, Beschreibung und konkrete textbasierte Belege aus dem Transkript dokumentiert. Die aus allen Segmenten extrahierten Symptome werden anschließend de-dupliziert und auf maximal fünf Kernsymptome verdichtet.

Der zweite Diagnose-Agent („diagnosis generator“) agiert als „...Experte für die Diagnoseerstellung nach ICD-10“ und generiert basierend auf der aggregierten Symptomliste eine strukturierte psychiatrische Beurteilung. Das System erstellt vollständige ICD-10-Diagnosen aus Kapitel V (F00-F99) mit entsprechenden Codes, Bezeichnungen und Schweregraden. Jede Diagnose wird durch eine systematische Überprüfung der Diagnosekriterien fachlich begründet und mit einem quantitativen Sicherheitswert (0-1 Skala) versehen. Bei komplexen Fällen können bis zu drei Diagnosen mit potenziellen Komorbiditäten erstellt werden.

Die durch *concisum* generierten Diagnosen, Zusammenfassungen und inhaltlichen Extraktionen wurden mit den bekannten Fallvignetten abgeglichen, um die Validität der Ergebnisse zu überprüfen. Aufgrund der begrenzten Stichprobengröße erfolgte keine statistische Auswertung; stattdessen wurde die Übereinstimmung im Rahmen eines qualitativ-deskriptiven Konsensverfahrens beurteilt.

## Ergebnisse

Unter Maßgabe der lokalen und datenschutzrechtlich unbedenklichen Auswertung vertraulicher Audioaufzeichnungen ließ sich eine handhabbare Pipeline für die Verarbeitung von Audiodateien bis zur Zusammenfassung von Therapiesitzungen sowie dem ersten Generieren von Verdachtsdiagnosen erstellen. Deren Erprobung anhand typischer Therapiegespräche resultierte in akkurate Transkripten mit einer guten Übereinstimmung der Zuordnung der Sprecher\*innen. Die erzielten Word Error Rates (WER) sowie die Word Diarization Error Rates (WDER) der automatisch erstellten Transkriptionen (im Vergleich zu den menschlichen Transkriptionen und Zuordnung der Sprecher\*innen) sind in Tabelle 1 zusammengefasst und zusätzlich in den Abbildungen 2 und 3 dargestellt. Die Transkriptionsgeschwindigkeit für *verbatim* (stereo und pyannote) lag in einem Echtzeitfaktor-Bereich von 0,51 bis 1,12, mit einem durchschnittlichen Echtzeitfaktor von 0,75 (d.h., einer Transkriptionsdauer von etwa 7,5 Minuten pro 10 Minuten Audiomaterial).

**Tabelle 1.**

### Transkriptionsgenauigkeit

Insgesamt ist die Transkriptionsgenauigkeit (Abbildung 2) bei der Verwendung von *verbatim*, ohne Normalisierung der Audiospuren, in nahezu allen untersuchten Fällen und gemittelt über alle Aufnahmemodalitäten mit einer mittleren Fehlerrate von 24,03 % niedriger als bei *openai/whisper* (mittlere Fehlerrate 31,13 %). Besonders deutlich fällt dieser Unterschied beim Premium-Funkmikrofon (Lavalier-Setup, PFM-L) aus: Hier erzielt *verbatim* mittlere WER-Werte zwischen 10,62 % (Spannweite: 7,88 – 13,35 %), während *openai/whisper* deutlich höhere Werte von durchschnittlich 21,71 % (Spannweite: 19,74 – 23,68 %) aufweist. Auch beim Premium-Funkmikrofon mit Tisch-Setup (PFM-T) liefert *verbatim* niedrigere mittlere Fehlerraten 19,46 % (Spannweite: 15,64 – 23,29 %) gegenüber *openai/whisper* 27,41 % (Spannweite 23,90 – 30,92 %). Beim Basismikrofon (BM) liegen die Fehlerraten hingegen insgesamt deutlich höher, jedoch mit ebenfalls niedrigeren Fehlerraten unter Verwendung von *verbatim*: 33,02 % (Spannweite: 25,12 – 40,10 %); *whisper*: 37,70 % (Spannweite: 30,01 – 51,51 %).

Auch die Art der Vorverarbeitung spielt eine relevante Rolle bei der erreichten Transkriptionsgenauigkeit: Über alle Setups hinweg liefert die *EBU R128*-Normalisierung die niedrigsten mittleren WER-Werte (*verbatim*: 21,74 %, *whisper*: 29,05 %), wobei die Unterschiede teilweise sehr klein ausfallen. *Speechnorm* (*verbatim*: 22,70 %, *whisper*: 29,05 %) führt in einigen Fällen ebenfalls zu einer Verbesserung gegenüber keiner Normalisierung, bleibt jedoch im Durchschnitt minimal hinter *EBU R128* zurück. Der Verzicht auf jegliche Normalisierung führt im Mittel zu den höchsten Fehlerraten (*verbatim*: 24,03 %, *whisper*: 31,13 %).

Insgesamt zeigt sich, dass *EBU R128* als Normalisierungsverfahren in nahezu allen getesteten Kombinationen zu niedrigeren WER führt – insbesondere in Verbindung mit *verbatim* und hochwertigen Mikrofon-Setups. Die *speechnorm*-Normalisierung bietet in manchen Fällen ebenfalls Verbesserungen, erreicht aber seltener die Qualität von *EBU R128*.

Auch in einer qualitativen, explorativen Betrachtung zeigen sich deutliche Unterschiede in der Transkriptionsqualität: Während bei den *whisper*-Transkriptionen immer wieder Konfabulationen auftreten (zum Beispiel eine 42-fache Wiederholung des Wortes „Ja“ in einer längeren Sprechpause) oder Einschübe von „garbled speech“ (wie „Ja, meinte er, fulfillment,'d fish like a light q'villa tree'r.“), lassen sich diese Fehltranskriptionen in *verbatim* zuverlässig unterdrücken.

### **Abbildung 2.**

#### **Zuordnung der Sprecher\*innen**

Ein differenzierteres Bild (siehe Abbildung 3) ergibt sich bei der (nicht normalisiert) Zuordnung der Sprecher\*innen (WDER). Durchschnittlich betrachtet, nur über Premium-Funkmikrofone hinweg, da es für die Basismikrofone keine Stereokanäle vorliegen, erzielt die Stereo-Analyse mit einer mittleren Fehlerrate von 9,19 % präzisere Ergebnisse als das Verfahren *pyannote* (Zuordnung mittels Deep-Learning) mit einer mittleren Fehlerrate von 13,93 %, allerdings variiert dies erheblich in Abhängigkeit der jeweiligen Aufnahmebedingungen. Besonders deutlich zeigt sich dieser Unterschied beim Lavalier-Setup (PFM-L), bei dem die Stereo-Analyse mit einer mittleren WDER von nur ca. 1,16 % (Spannweite: 0,40 – 1,91 %) deutlich präzisere Ergebnisse liefert als *pyannote* 23,91 % (Spannweite: 1,84 – 45,97 %). Auffällig ist hier insbesondere ein Ausreißer bei *pyannote*, welcher bei einer gleichgeschlechtlichen Dyade mit sehr ähnlichen Stimmen zu einer extrem hohen Fehlerrate von bis zu 45,97 % führte. Die Stereo-Analyse zeigte unter denselben Bedingungen hingegen maximal 1,91 % Fehler, was eine nahezu perfekte Sprecher\*innenuordnung bedeutet. Beim Premium-Funkmikrofon (Tisch-Setup, PFM-T) ist jedoch *pyannote* mit einer durchschnittlichen WDER von nur etwa 3,95 % (Spannweite: 2,60 – 5,29 %) präziser, während die Stereo-Analyse hier höhere Fehlerraten von durchschnittlich 17,22 % (Spannweite: ca. 9,60 – 24,76 %) aufweist.

Auch bei der Word Diarization Error Rate (WDER) zeigt sich ein klarer Einfluss der Vorverarbeitung: Die Methode *stereo* erzielt durchweg leicht niedrigere mittlere Fehlerraten als *pyannote*. Besonders in Kombination mit *EBU R128* erreicht *stereo* (gemittelt über Premium-Funk-Mikrofone-Setup) den besten Mittelwert (8,90 %), dicht gefolgt von keiner Normalisierung (9,19 %) und *speechnorm* (10,49 %). Bei *pyannote* verschlechtert eine Normalisierung mit *EBU R128* die Ergebnisse: Der mittlere WDER-Wert (gemittelt über Premium-Funk-Mikrofon-Setup) steigt unter *EBU R128* deutlich auf 23,55 %, während sowohl keine Normalisierung (13,93 %) als auch *speechnorm* (14,27 %) erheblich bessere Ergebnisse liefern. Insgesamt bestätigen die Ergebnisse, dass *stereo* die robustere Methode darstellt und *EBU R128* dort sinnvoll eingesetzt werden kann. Bei *pyannote* hingegen scheint eine Normalisierung nach *EBU R128* die Leistung deutlich zu verschlechtern. Hier liefern *speechnorm* oder der Verzicht auf Normalisierung die besseren Resultate.

### **Abbildung 3.**

Zusammenfassend lässt sich festhalten, dass *verbatim* grundsätzlich genauere Transkriptionen liefert und die Stereo-Analyse in vielen, jedoch nicht allen Fällen, Vorteile bei der Zuordnung der Sprecher\*innen bietet. Insbesondere beim Lavalier-Setup ist die Stereo-Analyse klar präziser, während beim Tischmikrofon *pyannote* überlegen ist. Das Basismikrofon eignet sich hingegen insgesamt kaum für eine zuverlässige automatische Transkription.

#### **Zusammenfassungen und Verdachtsdiagnosen**

Drei der vier generierten Zusammenfassungen und Verdachtsdiagnosen sowie eine Kritische Würdigung der produzierten Resultate finden sich in Tabelle 2 (die vierte in Appendix B). Insgesamt erscheinen die generierten Zusammenfassungen valide und die wesentlichen Eckdaten des Gesprächs sind enthalten und gut erfasst. Die abgeleiteten Verdachtsdiagnosen sind nahe an den von den Schauspielpatient\*innen dargestellten Krankheitsbildern, manche zusätzlichen Verdachtsdiagnosen erscheinen jedoch eher spekulativ. Bei der Begründung der Diagnosestellung und der Belegführung, ergeben sich jedoch einige gravierende Fehler. So werden beispielsweise relevante Zitate aus dem Gespräch aufgeführt, diese aber teilweise der falschen Person zugeschrieben, zum Beispiel: „[Therapeut]: Ich trinke schon jeden Tag.“, in einer Vignette, in welcher der Patient kritischen Alkoholkonsum einräumt.

Ein weiterer augenfälliger Fehler ist, dass falsche Anglizismen in die Texte einfließen, z.B. „delusionale Störung“ statt „wahnhafte Störung“.

## Tabelle 2.

### Diskussion

Im Zentrum der Psychotherapie steht der sprachliche Austausch, welcher routinemäßig zu dokumentieren ist, aber aufgrund seiner Komplexität sowie dem damit verbundenen Zeitaufwand kaum für die direkte Analyse zur Verfügung steht. In der vorliegenden pilotierenden Machbarkeitsstudie wurde die Standardimplementierung von OpenAI mit einer angepassten und weiterentwickelten, innovativen Methode verglichen. Die Ergebnisse verdeutlichen die Leistungsfähigkeit aktueller Transkriptionsmodelle im Kontext der Psychotherapie, wobei ein besonderer Fokus auf die Qualitätssicherung (Reduktion von Konfabulationen sowie der Verbesserung der Zuordnung der Sprecher\*innen) gelegt wurde und gleichzeitig mit einer vollständig lokalen Offline-Anwendung gearbeitet wurde. Im Rahmen dieser Pilotierung konnte gezeigt werden, dass moderne Transkriptionssysteme, selbst unter ressourcenoptimierten Bedingungen und unter Verwendung handelsüblicher Technik, ein sehr großes Potenzial für die psychotherapeutische Praxis und Wissenschaft bieten können.

In diesem spezifischen Setting konnte gezeigt werden, dass die hier verwendeten lokalen Methoden (*verbatim*) bezüglich der Word-Error-Rate (WER) bessere Ergebnisse lieferten als OpenAIs Standard-Implementierung (*openai/whisper*). Die erzielten Ergebnisse deuten darauf hin, dass die Verwendung von *verbatim* hinsichtlich der Word Error Rate (WER) über alle vier Dyaden zu besseren Ergebnissen führen. Besonders deutlich wird dieser Unterschied beim Einsatz eines Premium-Funkmikrofons im Lavalier-Setup, bei dem *verbatim* durchschnittlich ca. 10 % WER erreichte, gegenüber ca. 23 % bei der Standard-Implementierung. Den niedrigeren Fehlerraten stehen im Zusammenhang mit einer Verringerung der Konfabulationen, welche insbesondere im medizinischen und psychotherapeutischen Kontext von hoher Bedeutung sein können, da erfundene oder verfälschte Worte die therapeutischen Zusammenfassungen und weiteren Analysen gefährlich verändern können [Koenecke et al., 2024].

Die hier eingesetzte OpenAI Standard-Implementierung (*openai/whisper*) erreichte nur etwas geringere WER als in anderen Evaluationsstudien unter anderen Rahmenbedingungen (z. B. [Radford et al., 2022]). Ein Grund liegt wahrscheinlich darin, dass *openai/whisper* besonders gut geeignet ist für Datensätze, die dem Trainings-Corpus ähnlich sind (z. B. englischsprachige Audios). Die Anwendung in einem therapeutischen Setting – welches sich signifikant von klassischen Evaluationsbedingungen unterscheidet – dürfte deshalb spezielle Anpassungen erfordern (wie durch *verbatim* teilweise umgesetzt), damit eine adäquate Leistungsfähigkeit im veränderten Anwendungskontext gewährleistet ist. Genau dies unterstreicht das Potential von *verbatim* in diesem Setting. Dennoch sind selbstverständlich weitere Validierungen mit größeren Stichproben erforderlich.

Die akkurate Zuordnung der Sprecher\*innen stellte eine weitere zentrale Leistung dieser Demonstration dar. Hier zeigten die Ergebnisse, dass der Einsatz von Stereo-Aufnahmen mit getrennter Signalverarbeitung generell zu einer verbesserten Zuordnung der Sprecher\*innen führte (Stereo mittlere WDER = 9,2 % vs. *pyannote* mittlere WDER = 17,2 %). Besonders eindrucksvoll war die Verbesserung bei einer gleichgeschlechtlichen Dyade mit äußerst ähnlichen Stimmen, bei welcher durch die Stereo-Analyse im Lavalier-Setup eine drastische Reduktion der WDER von maximal 45,97 % (*pyannote*) auf nur 1,91 % erreicht wurde. Dieses Resultat deutet darauf hin, dass Stereo-Analysen insbesondere bei zum Beispiel ähnlichen Stimmen sehr nützlich sein können. Jedoch lässt sich dies nur dann gewinnbringend einbringen, wenn die jeweils ein\*e Sprecher\*in deutlich lauter auf einem Kanal aufgezeichnet wird. So zeigte sich, dass in spezifischen Fällen, insbesondere beim Tisch-Setup mit klar differenzierbaren Stimmen, *pyannote* mit einer niedrigeren WDER (durchschnittlich 3,95 %) bessere Ergebnisse erzielte als die Stereo-Analyse (durchschnittlich 17,22 %). Damit wird deutlich, dass der Nutzen der Stereo-Technologie insbesondere bei großer Ähnlichkeit der Sprecher\*innenstimmen hilfreich sein kann, sofern sich die physikalische Lautstärke zwischen den Audiospuren ausreichend unterscheidet. Generell kann daher eine Stereoaufzeichnung empfohlen werden, die je nach Dyade für beide Auswertungswege verwendet werden kann.

Die Wahl des Normalisierungsverfahrens (*EBU R128*, „*speechnorm*“ oder keine Normalisierung) beeinflusste sowohl die WER als auch die WDER, allerdings differenziert und teilweise gegensätzlich. Während *EBU R128* (insbesondere bei hochwertigen Mikrofonen) sowohl unter *openai/whisper* wie auch innerhalb von *verbatim* in Verbindung mit der Stereo-Methode sowohl die Transkriptions- als auch die Qualität der Zuordnung der Sprecher\*innen verbesserte, führte dasselbe Verfahren bei *pyannote* zu

erhöhten Fehlerraten. Die Ursachen hierfür bleiben vorerst unklar und könnten auf subtile, möglicherweise ungünstige Veränderungen der relevanten akustischen Merkmale zurückzuführen sein. Um diese Effekte genauer zu verstehen, sind weitere Untersuchungen, insbesondere mit einem größeren Datensätzen, notwendig. Dennoch empfiehlt sich bei *pyannote* der Verzicht auf eine Normalisierung oder der Einsatz alternativer Verfahren wie „*speechnorm*“. Vertiefende Forschung sollte diese Interaktion genauer untersuchen, um methodenspezifische Vorverarbeitungsstrategien klarer ableiten zu können.

Ein weiterer wichtiger Erkenntnisgewinn der Untersuchung war die Fähigkeit, automatisch zusammengefasste Transkripte sowie erste diagnostische Vorschläge zu generieren. Die generierten Zusammenfassungen und diagnostischen Ansätze, die mithilfe von *concisum* erstellt wurden, wiesen eine hohe inhaltliche Kohärenz und Relevanz auf. Die vorgeschlagenen Verdachtsdiagnosen lagen fachlich nahe an den ursprünglich intendierten Diagnosen der Schauspielpatient\*innen; beispielsweise wurden bei einer intendierten Alkoholabhängigkeit ein schädlicher Gebrauch und Abhängigkeit bezüglich der Codierung durchmischt. Ob dies auf die eingeschränkte Informationsbasis des simulierten Erstgesprächs oder auf mangelnde diagnostische Differenzierungsfähigkeit der eingesetzten Modelle zurückzuführen ist, lässt sich anhand der vorliegenden Daten nicht abschließend beurteilen und sollte in zukünftigen Studien weiter untersucht werden. Die in dieser Pilotstudie eingesetzte mehrstufige, hierarchische Zusammenfassungs-Pipeline erhöht die Transparenz diagnostischer Prozesse und ermöglicht eine nachvollziehbare Herleitung der vom System vorgeschlagenen Diagnosen. Einfachere Prompts (siehe Appendix C) führen zwar zu ähnlich guten Zusammenfassungen und plausibel erscheinenden Verdachtsdiagnosen, bieten jedoch deutlich weniger Einblick in die zugrundeliegenden diagnostischen Entscheidungen. Daher stellt der hier verwendete System *concisum* einen vielversprechenden Ansatz für zukünftige Entwicklungen dar. Allerdings offenbaren die in dieser Untersuchung beobachteten Fehler bei der Herleitung einzelner Diagnosen deutliche Grenzen hinsichtlich einer derzeitigen Anwendung in der klinischen Praxis: Die Ergebnisse vermischen häufig Symptom-, Beschreibungs- und Diagnoseebenen, wie bei der fehlerhaften Zuordnung von Alkoholabhängigkeit als Symptom, und lassen zentrale diagnostische Kriterien unberücksichtigt. Obwohl die Vorschläge thematisch oft nah an den intendierten Störungsbildern liegen, führen unscharfe Terminologie, lückenhafte Begründungen und falsche Klassifikationen zu unpräzisen Ergebnissen. Zukünftige Forschung sollte verstärkt Retrieval-Augmented-Generation-Ansätze integrieren, um die LLM-basierte Textgeneration durch optimierte Kontextdarbietung relevanter Fachliteratur fachlich korrekter und nachvollziehbarer zu machen.

Insgesamt zeigen die hier beschriebenen explorativen Ergebnisse, dass große Sprachmodelle, die lokal betrieben werden können, eine vielversprechende Unterstützung für die Automatisierung administrativer Aufgaben in der Psychotherapie darstellen können. Durch die Entlastung der Therapeut\*innen von zeitaufwändigen Aufgaben wie der Protokollierung und Berichterstellung könnten wertvolle Ressourcen für die direkte therapeutische Arbeit freigesetzt werden, was in Anbetracht des steigenden administrativen Drucks und der begrenzten Therapiezeit eine wichtige Entwicklung darstellt [Henning et al., 2021; Piechotta, 2008].

In der Psychotherapie wird mit hochsensiblen Daten gearbeitet [Alpers et al., 2020]. Die Möglichkeit, KI-Modelle lokal und sogar getrennt vom Netzwerk (air-gapped) zu nutzen, eröffnet neue Handlungsspielräume im Umgang mit dem Zielkonflikt zwischen Skalierbarkeit, Nutzerfreundlichkeit, Verfügbarkeit einerseits sowie Datenhoheit und Sicherheit andererseits. Außerdem könnte dies auch dabei helfen, das Vertrauen der Patient\*innen in den sicheren Umgang mit ihren Daten zu stärken (wo dies gerechtfertigt ist). Trotz dieser Fortschritte bedarf es weiterhin intensiver Forschungsbemühungen, um die langfristige Sicherheit und Verlässlichkeit dieser Systeme, insbesondere in Bezug auf die Speicherung und Nutzung sensibler Daten, zu gewährleisten.

Bei weiterer Dokumentation der Zuverlässigkeit und niederschwelliger(er) technischer Anwendung lassen sich solche Methoden für Gespräche in der Praxis wie für Interviews zu Forschungszwecken in allen möglichen Wissenschaftsdisziplinen verwenden, auch, um qualitative Analysen weiterzuentwickeln. Im konkreten Fall der Prädiktion relevanter therapeutischer Outcomes [vergl. Taubitz et al., 2022] stehen also künftig weitere Datenquellen neben den meist auf Selbstbericht basierenden psychometrischen Daten zur Verfügung. Wird angestrebt, Large Language Models zur Unterstützung von Interventionen heranzuziehen [Alpers, 2023], ist es unerlässlich, die in der Therapie gesprochenen Inhalte zu verschriftlichen. Darüber hinaus werden weitere Quellen, die zusätzliche Ebenen der emotionalen Prozesse im therapeutischen

Prozess erfassen können, hinzukommen, beispielsweise KI-basierte Analysen des emotionalen Gesichtsausdrucks [Höfling und Alpers, 2023].

Trotz dieser vielversprechenden Illustrationen sind relevante Limitationen zu beachten. So ist die begrenzte Materialvielfalt unserer Pilottestung sowie die Verwendung von simulierten Erstgesprächen ein Faktor, der die Generalisierbarkeit der Ergebnisse einschränkt. Auch wenn der naturalistische Ansatz und die Wahl handelsüblicher Technik die Relevanz der Ergebnisse für den praktischen Alltag erhöhen, bleibt offen, wie sich diese Systeme in einem breiteren Spektrum von realen Anwendungsszenarien bewähren, etwa bei komplexeren oder länger dauernden Therapiegesprächen.

Weiterhin lässt sich, bei der die hier verwendeten kleinen Datenbasis, keine eindeutige (inferenzstatistische) Überlegenheit bezüglich der Wahl einer Methode zur Sprecher\*innenzuweisung ableiten. Zwar wird derzeit in *verbatim* an der Implementierung eines automatischen Algorithmus gearbeitet, der in Abhängigkeit vom Setup und den Eigenschaften der Sprecher\*innen eine optimale Zuordnungsmethode vorschlägt, um den Entscheidungsprozess zu vereinfachen; jedoch sollte die Zuverlässigkeit dieser Methode auch anhand umfangreicherer Testdaten, nicht zuletzt inferenzstatistisch evaluiert werden. Zudem erfordern Sprecher\*innen, die sich im Raum bewegen (z. B. in kinder- und jugendpsychotherapeutischen Kontexten) wahrscheinlich spezifische Anpassungen, um etwaigen Einbußen der Aufnahmegeradeität zu kompensieren. Zukünftige Studien sollten daher noch mehr Varianten von Audiodateien, unter anderem auch eine Dyade männlicher Sprecher, sowie diverse therapeutische Settings einbeziehen, um die Robustheit und Vielseitigkeit der Modelle weiter zu validieren. Die zu diesem Zweck händische Transkription ist jedoch aufwändig und zeitintensiv, soll sie so akribisch erfolgen wie die hier als Ground-Truth verwendeten Transkriptionen.

Darüber hinaus scheint der Vergleich mit weiteren State-of-the-Art Algorithmen, wie zum Beispiel WhisperX [Bain et al., 2023] zukünftig sinnvoll. Weil wir angestrebt haben, konventionelle Hardware zu erproben, war dies in dieser Studie nicht möglich.

Auch die Wahl der verwendeten Mikrofone muss im jeweiligen Anwendungsbereich abgewogen werden. Möglicherweise sind Ansteckmikrofone besonders salient, was sich auf das Therapiegeschehen auswirken könnte. Wir konnten jedoch nicht beobachten, dass Ansteckmikrofone eine höhere Salienz haben als ebenfalls sichtbare Tischmikrofone. Offenbar berichteten aber in anderen Studien einzelne Personen, dass sie sich angesichts von Tonaufnahmen im Gespräch unwohl fühlen, es gibt aber keine Belege für einen bedeutsamen Einfluss auf das Therapieergebnis [Briggle et al., 2016].

Selbstverständlich entscheidet die aktuelle technische Ausstattung über künftige Verwendungsmöglichkeiten. Sollen archivierte Daten z.B. einer Hochschulambulanz ausgewertet werden, stehen bisher kaum Stereoaufnahmen zur Verfügung. Um diese Daten künftig dennoch für Forschungszwecke nutzbar zu machen, könnten andere Ansätze hilfreich sein, mit denen einzelne Sprecher\*innen auch in Mono-Aufnahmen zuverlässig identifiziert werden können. So wäre es denkbar, individuelle Sprachprofile zu erstellen und zu differenzieren, um Therapeuten und Patienten in der Aufnahme automatisch unterscheiden zu können [z.B. mittels SpeechBrain; vgl. Dawalatabad et al., 2021]. Wir haben diesen Ansatz nicht gewählt, weil dabei vorab geeignete Abschnitte identifiziert werden müssten, in denen jeweils nur eine Person spricht bzw. explizit eine Sprachprobe abgibt, was einen Eingriff in die Therapiesituation darstellen würde. Auch eine Zuordnung der Sprecher\*innen anhand inhaltlicher Merkmale wäre möglich, erscheint uns aber angesichts häufiger Paraphrasierungen in der Psychotherapie besonders fehleranfällig zu sein.

Das vorgestellte Verfahren ermöglicht die datenschutzkonforme Auswertung von Therapiegesprächen unter vollständiger Datenhoheit. Dennoch bleiben eine sachgerechte Implementierung sowie eine fallspezifische technische und rechtliche Prüfung unerlässlich, um den angemessenen Einsatz dieses oder vergleichbarer Methoden zu gewährleisten.

Obwohl dies keine Einschränkung der Studie an sich darstellt, sollen weitere kritische Implikationen reflektiert werden: Trotz vielversprechender Ergebnisse, plädieren wir dafür die erstellten Transkripte und Zusammenfassungen stets zu überwachen. Die Automatisierung eines lästigen Prozesses darf nicht dazu führen, dass wichtige Informationen übersehen werden, die bei einer gründlichen händischen Beschäftigung mit der Protokollierung eventuell Beachtung gefunden hätten. Es ist nicht auszuschließen, dass gerade der mühselige Vorgang der Protokollierung hilft, die Informationen noch einmal in einem anderen/größeren Zusammenhang zu betrachten. So wurde verschiedentlich auf die strukturierende Funktion des kognitiven Prozesses beim aktiven Schreiben hingewiesen [Oatley und Djikic, 2008] und die Transkription als mehr als

ein handwerklicher Prozess erfasst [Braun und Clarke, 2006]. Dies ist insbesondere bei den interpretierenden (bzw. diagnostizierenden) Zusammenfassungen relevant, da es hier nicht nur um die korrekte Protokollierung, sondern vielmehr um die informierte Interpretation des therapeutischen Prozesses ankommt. Beispielsweise könnten sich Beiträge zum therapeutischen Gespräch widersprechen und die letztendlich getroffene Formulierung damit für den therapeutischen Ausblick schwerer wiegen als ein anfängliches Gedankenspiel.

Insgesamt legen diese ersten Ergebnisse nahe, dass KI-gestützte Systeme ein großes Potenzial zur Verbesserung der Transkription und Zuordnung der Sprecher\*innen im psychotherapeutischen Kontext besitzen. Insbesondere die Verbesserung der Transkription und die ressourcenoptimierte Anpassung der Zuordnung der Sprecher\*innendurch den Einsatz von Stereo-Technologie stellen bedeutende Fortschritte dar, von denen auch andere Disziplinen, die mit Interviews arbeiten, profitieren können. Darüber hinaus, ist insbesondere die Verwendung handelsüblicher Hardware und frei verfügbarer und unlizenzierte Software von elementarer Bedeutung, da dies die Praxisrelevanz unserer Ergebnisse obendrein unterstreicht. Jedoch bleibt die Notwendigkeit, die Systeme weiter zu optimieren und zu validieren, um deren Anwendung in verschiedenen therapeutischen Kontexten und bei realen Patient\*innen zu gewährleisten. Zukünftige Forschung sollte sich auf größere Datenvolumina und die Integration der Modelle in komplexe, längere Therapiegespräche konzentrieren, um ihre Effektivität und Praxistauglichkeit weiter zu evaluieren. Nur so kann sichergestellt werden, dass KI-Technologien die Qualität der psychotherapeutischen Versorgung langfristig steigern und gleichzeitig die ethischen und datenschutzrechtlichen Anforderungen erfüllen.

## Acknowledgements

Wir danken Maria Guelguisser, B.Sc, für die Unterstützung bei den manuellen Transkriptionen, sowie den Studierenden und Schauspielpatient\*innen für die Teilnahme. Wir danken Tobias Morweiser für umsichtigen Techniksupport. Im Schreibprozess verwendeten die Autoren ChatGPT [OpenAI, 2025] sowie Grammarly [Grammarly Inc., 2025], um die Lesbarkeit zu verbessern. Die Autoren haben stets den Inhalt überprüft und bearbeitet und übernehmen die volle Verantwortung für den Inhalt des veröffentlichten Artikels.

## Statement of Ethics

Alle Teilnehmenden erteilten nach Aufklärung ihre schriftliche, informierte Einwilligung. Eine Genehmigung durch die Ethikkommission war gemäß den lokalen und nationalen Richtlinien nicht erforderlich. Die Autoren bestätigen die Einhaltung der aktuellen Fassung der Deklaration von Helsinki.

## Conflict of Interest Statement

Alle Autoren erklären, dass kein Interessenskonflikt besteht. Sie haben nicht die Absicht, für kommerzielle Produkte oder Angebote zu werben.

## Funding Sources

Das diesem Bericht zugrunde liegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung (BMBF) unter dem Förderkennzeichen 03IHS264B im Rahmen von TransforMA gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autoren.

## Author Contributions

**Friedrich-Samuel Taubitz:** Konzeption; Methodik; Untersuchung; Datenerhebung; Datenaufbereitung; Datenkurierung; Softwarekonzeptualisierung; formale Analyse; Validierung; Verfassen des Originalentwurfs; Überarbeitung und Redaktion. **Linus H. A. Sehn:** Methodik; Untersuchung; Datenaufbereitung; Datenkurierung; Software; formale Analyse; Validierung; Überarbeitung und Redaktion. **Georg W. Alpers:** Konzeption; Methodik; Untersuchung; Ressourcenbereitstellung; Überarbeitung und Redaktion; Supervision; Projektadministration (Projektleitung, Principal Investigator). Alle Autoren diskutierten die Ergebnisse, überarbeiteten das Manuskript gemeinsam und genehmigten die endgültige Fassung.

## Data Availability Statement

Analyseskripte und Methoden sind auf GitHub frei zugänglich [Petit und Sehn, 2025; Sehn, 2025]; die zugrunde liegenden Daten werden aus Datenschutzgründen nicht öffentlich bereitgestellt, können aber bei berechtigtem wissenschaftlichem Interesse auf Anfrage an die korrespondierenden Autor\*innen eingesehen werden.

## Literaturverzeichnis

- Alpers GW. Wird ChatGPT auch in der Psychotherapie Einzug halten? Ein Expert\*inneninterview. Psychotherapeutenjournal. 2023; 22:156–160.
- Alpers GW, Frey L, Tessmer-Petzendorfer S, Klingauf A, und Schad S. Wie schreib ich's meinem Patienten? Datenschutzprobleme bei der digitalen Kommunikation. Verhaltenstherapie. 2020; 30:334–346.  
<https://doi.org/10.1159/000504147>
- Alpers GW und Hengen KM. Interactions with Standardized Patients to Evaluate Students' Psychotherapy-Competencies: Reliable Assessment and Valid Evaluation. Z Für Klin Psychol Psychother. 2021; 50:133–144.  
<https://doi.org/10.1026/1616-3443/a000636>
- Alpers GW und Steiger-White F. Simulationspatienten im universitären Psychologiestudium: Einführung einer innovativen Methode zur Stärkung praxisorientierter Lehre in klinischer Psychologie und Psychotherapie. Verhaltenstherapie. 2020; 30:104–116. <https://doi.org/10.1159/000507740>
- Alpers GW, Winzelberg AJ, Classen C, Roberts H, Dev P, Koopman C, et al. Evaluation of computerized text analysis in an Internet breast cancer support group. Comput Hum Behav. 2005; 21:361–376.  
<https://doi.org/10.1016/j.chb.2004.02.008>
- Bain M, Huh J, Han T, und Zisserman A. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio [Internet]. 2023. <https://doi.org/10.48550/arXiv.2303.00747>
- Braun V und Clarke V. Using thematic analysis in psychology. Qual Res Psychol. 2006; 3:77–101.  
<https://doi.org/10.1191/1478088706qp063oa>
- Bredin H. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe [Internet]. In: Interspeech 2023; Dublin, Ireland. ISCA; 2023. pp 1983–1987. <https://doi.org/10.21437/Interspeech.2023-105>
- Briggie AM, Hilsenroth MJ, Conway F, Muran JC, und Jackson JM. Patient comfort with audio or video recording of their psychotherapy sessions: Relation to symptomatology, treatment refusal, duration, and outcome. Prof Psychol Res Pract. 2016; 47:66–76. <https://doi.org/10.1037/a0040063>
- Buchholz MB und Kächele H. From turn-by-turn to larger chunks of talk: An exploratory study in psychotherapeutic micro-processes using conversation analysis. Res Psychother Psychopathol Process Outcome. 2017; 20. <https://doi.org/10.4081/ripppo.2017.257>
- Cummins R, Ewbank MP, Martin A, Tablan V, Catarino A, und Blackwell AD. TIM: A Tool for Gaining Insights into Psychotherapy [Internet]. In: WWW '19: The Web Conference; San Francisco CA USA. ACM; 2019. pp 3503–3506. <https://doi.org/10.1145/3308558.3314128>
- Dawalatabad N, Ravanelli M, Grondin F, Thienpondt J, Desplanques B, und Na H. ECAPA-TDNN Embeddings for Speaker Diarization [Internet]. In: Interspeech 2021. ISCA; 2021. pp 3560–3564.  
<https://doi.org/10.21437/Interspeech.2021-941>
- Défossez A, Usunier N, Bottou L, und Bach F. Demucs: Deep Extractor for Music Sources with extra unlabeled data remixed [Internet]. 2019. <https://doi.org/10.48550/arXiv.1909.01174>
- Eberhardt ST, Schaffrath J, Moggia D, Schwartz B, Jaehde M, Rubel JA, et al. Decoding emotions: Exploring the validity of sentiment analysis in psychotherapy. Psychother Res. 2024; 35:174–189.  
<https://doi.org/10.1080/10503307.2024.2322522>
- European Broadcasting Union. R128: Loudness Normalisation and Permitted Maximum Level of Audio Signals [Internet]. 2023 [zugegriffen 10 Apr. 2025]. Verfügbar unter: <https://tech.ebu.ch/docs/r/r128.pdf>
- Ewbank MP, Cummins R, Tablan V, Bateup S, Catarino A, Martin AJ, et al. Quantifying the Association Between Psychotherapy Content and Clinical Outcomes Using Deep Learning. JAMA Psychiatry. 2019; 77:35–43. <https://doi.org/10.1001/jamapsychiatry.2019.2664>
- Ewbank MP, Cummins R, Tablan V, Catarino A, Buchholz S, und Blackwell AD. Understanding the relationship between patient language and outcomes in internet-enabled cognitive behavioural therapy: A deep learning approach to automatic coding of session transcripts. Psychother Res. 2021; 31:300–312.  
<https://doi.org/10.1080/10503307.2020.1788740>
- FFmpeg Developers. FFmpeg/FFmpeg [Internet]. 2025 [zugegriffen 10 Apr. 2025]. Verfügbar unter: <https://github.com/FFmpeg/FFmpeg>

- Fraunfelter L-A, Gerdes ABM, und Alpers GW. Spieglein, Spieglein an der Wand: Verzerrte Selbstbewertungen psychotherapeutischer Kompetenzen im Psychotherapie-Studiengang. *PPmP - Psychother · Psychosom · Med Psychol.* 2025; 75:162–172. <https://doi.org/10.1055/a-2547-8479>
- Grammarly Inc. Grammarly [Internet]. 2025 [zugegriffen 11 Mai 2025]. Verfügbar unter: <https://www.grammarly.com>
- Henning A, Helmbold N, und Diel F. Sektorenübergreifende Qualitätssicherung – Impulse für eine Neuausrichtung. *Psychother Aktuell.* 2021; Sonderausgabe 2021:28–33.
- Höfling TTA und Alpers GW. Automatic facial coding predicts self-report of emotion, advertisement and brand effects elicited by video commercials. *Front Neurosci.* 2023; 17. <https://doi.org/10.3389/fnins.2023.1125983>
- Hur JK, Heffner J, Feng GW, Joormann J, und Rutledge RB. Language sentiment predicts changes in depressive symptoms. *Proc Natl Acad Sci.* 2024; 121:e2321321121. <https://doi.org/10.1073/pnas.2321321121>
- Koenecke A, Choi ASG, Mei KX, Schellmann H, und Sloane M. Careless Whisper: Speech-to-Text Hallucination Harms [Internet]. In: FAccT '24: The 2024 ACM Conference on Fairness, Accountability, and Transparency; Rio de Janeiro Brazil. ACM; 2024. pp 1672–1681. <https://doi.org/10.1145/3630106.3658996>
- Lalk C, Steinbrenner T, Kania W, Popko A, Wester R, Schaffrath J, et al. Measuring Alliance and Symptom Severity in Psychotherapy Transcripts Using Bert Topic Modeling. *Adm Policy Ment Health Ment Health Serv Res.* 2024; 51:509–524. <https://doi.org/10.1007/s10488-024-01356-4>
- Lasswell H. Verbal references and physiological changes during the psychoanalytic interview: a preliminary communication. *Psychoanal Rev.* 1935; :10–24.
- Lin B, Cecchi G, und Bouneffouf D. Working Alliance Transformer for Psychotherapy Dialogue Classification [Internet]. 2022a [zugegriffen 6 Mai 2024]. Verfügbar unter: <http://arxiv.org/abs/2210.15603>
- Lin B, Cecchi G, und Bouneffouf D. SupervisorBot: NLP-Annotated Real-Time Recommendations of Psychotherapy Treatment Strategies with Deep Reinforcement Learning [Internet]. 2022b. <https://doi.org/10.48550/arXiv.2208.13077>
- Lutz W, Schaffrath J, Eberhardt ST, Hehlmann MI, Schwartz B, Deisenhofer A-K, et al. Precision Mental Health and Data-Informed Decision Support in Psychological Therapy: An Example. *Adm Policy Ment Health.* 2023; 51:674. <https://doi.org/10.1007/s10488-023-01330-6>
- Lutz W, Vehlen A, und Schwartz B. Data-informed psychological therapy, measurement-based care, and precision mental health. *J Consult Clin Psychol.* 2024; 92:671–673. <https://doi.org/10.1037/ccp0000904>
- Miner AS, Haque A, Fries JA, Fleming SL, Wilfley DE, Terence Wilson G, et al. Assessing the accuracy of automatic speech recognition for psychotherapy. *Npj Digit Med.* 2020; 3:1–8. <https://doi.org/10.1038/s41746-020-0285-8>
- Naffah A, Pfeifer VA, und Mehl MR. Spoken Language Analysis in Aging Research: The Validity of AI-Generated Speech to Text Using OpenAI's Whisper. *Gerontology.* 2025; 71:417–424. <https://doi.org/10.1159/000545244>
- Oatley K und Djikic M. Writing as Thinking. *Rev Gen Psychol.* 2008; 12:9–27. <https://doi.org/10.1037/1089-2680.12.1.9>
- OpenAI. ChatGPT [Internet]. 2025 [zugegriffen 3 Mai 2025]. Verfügbar unter: URL = <https://chat.openai.com>
- Petit G und Sehn L. gaspardpetit/verbatim [Internet]. 2025 [zugegriffen 15 Jan. 2025]. Verfügbar unter: <https://github.com/gaspardpetit/verbatim>
- Piechotta B. Qualitätsmanagement in der ambulanten Psychotherapie [Internet]. In: ; Berlin, Heidelberg. Springer; 2008. pp 43–100. [https://doi.org/10.1007/978-3-540-75742-9\\_2](https://doi.org/10.1007/978-3-540-75742-9_2)
- Plaquet A und Bredin H. Powerset multi-class cross entropy loss for neural speaker diarization [Internet]. In: INTERSPEECH 2023; Dublin, Ireland. 2023. pp 3222–3226. <https://doi.org/10.21437/Interspeech.2023-205>
- Radford A, Kim JW, Xu T, Brockman G, McLeavey C, und Sutskever I. Robust Speech Recognition via Large-Scale Weak Supervision [Internet]. 2022. <https://doi.org/10.48550/ARXIV.2212.04356>
- Rogers CR. The use of electrically recorded interviews in improving psychotherapeutic techniques.. *Am J Orthopsychiatry.* 1942; 12:429–434. <https://doi.org/10.1111/j.1939-0025.1942.tb05930.x>
- Sehn L. linozen/concisuim [Internet]. 2025 [zugegriffen 17 Juli 2025]. Verfügbar unter: <https://github.com/linozen/concisuim>
- Siddharth SN und Beigi H. Bi-LSTM Scoring Based Similarity Measurement with Agglomerative Hierarchical Clustering (AHC) for Speaker Diarization. 2022. <https://doi.org/10.13140/RG.2.2.13977.29288>

Silero Team. Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier [Internet]. 2024 [zugegriffen 3 Apr. 2025]. Verfügbar unter:  
<https://github.com/snakers4/silero-vad>

Singer S, Maier L, Paserat A, Lang K, Wirp B, Kobes J, et al. Wartezeiten auf einen Psychotherapieplatz vor und nach der Psychotherapiestrukturreform. *Psychotherapeut*. 2022; 67:176–184.

<https://doi.org/10.1007/s00278-021-00551-0>

Smith AL, Greaves F, und Panch T. Hallucination or Confabulation? Neuroanatomy as metaphor in Large Language Models. *PLOS Digit Health*. 2023; 2:e0000388. <https://doi.org/10.1371/journal.pdig.0000388>

Strömsdörfer H. Praxen an der Belastungsgrenze – ambulante Versorgung stärken! [Internet], Berlin, Deutsche Psychotherapeuten Vereinigung, 2023, [zugegriffen 15 Jan. 2025].

Taubitz F-S, Büdenbender B, und Alpers GW. What the future holds: Machine learning to predict success in psychotherapy. *Behav Res Ther*. 2022; 156:104116. <https://doi.org/10.1016/j.brat.2022.104116>

Wang Q, Huang Y, Zhao G, Clark E, Xia W, und Liao H. DiarizationLM: Speaker Diarization Post-Processing with Large Language Models [Internet]. 2024. <https://doi.org/10.48550/arXiv.2401.03506>

Wittberg V, Kluge H-G, Rottmann H, und Goldmann E. Der Bürokratieindex für die vertragsärztliche Versorgung. Belastung transparent machen, Bürokratie abbauen [Internet]. Kassenärztliche Bundesver. 2020 [zugegriffen 21 März 2025]. Verfügbar unter: [https://www.kbv.de/media/sp/BIX2020\\_Projektbericht.pdf](https://www.kbv.de/media/sp/BIX2020_Projektbericht.pdf)

Zhao S, Ma B, Watcharasupat KN, und Gan W-S. FRCRN: Boosting Feature Representation using Frequency Recurrence for Monaural Speech Enhancement [Internet]. 2024. <https://doi.org/10.48550/arXiv.2206.07293>

Zhou L, Blackley SV, Kowalski L, Doan R, Acker WW, Landman AB, et al. Analysis of Errors in Dictated Clinical Documents Assisted by Speech Recognition Software and Professional Transcriptionists. *JAMA Netw Open*. 2018; 1:e180530. <https://doi.org/10.1001/jamanetworkopen.2018.0530>

## Abbildungen Legenden

### Abbildung 1. Schematischer Überblick der Prozessschritte

Anmerkung: Die Darstellung ist bewusst stark vereinfacht; einzelne technische Details wurden zur Übersichtlichkeit ausgelassen, sodass geringfügige Ungenauigkeiten möglich sind.

### Abbildung 2. Vergleich der Word Error Rate (WER) bei vier Sprecher\*innen-Paaren.

Verglichen werden eine etablierte Methode (*openai/whisper*) und *verbatim* unter unterschiedlichen Normalisierungen, Mikrofontypen und Mikrofon-Setups.

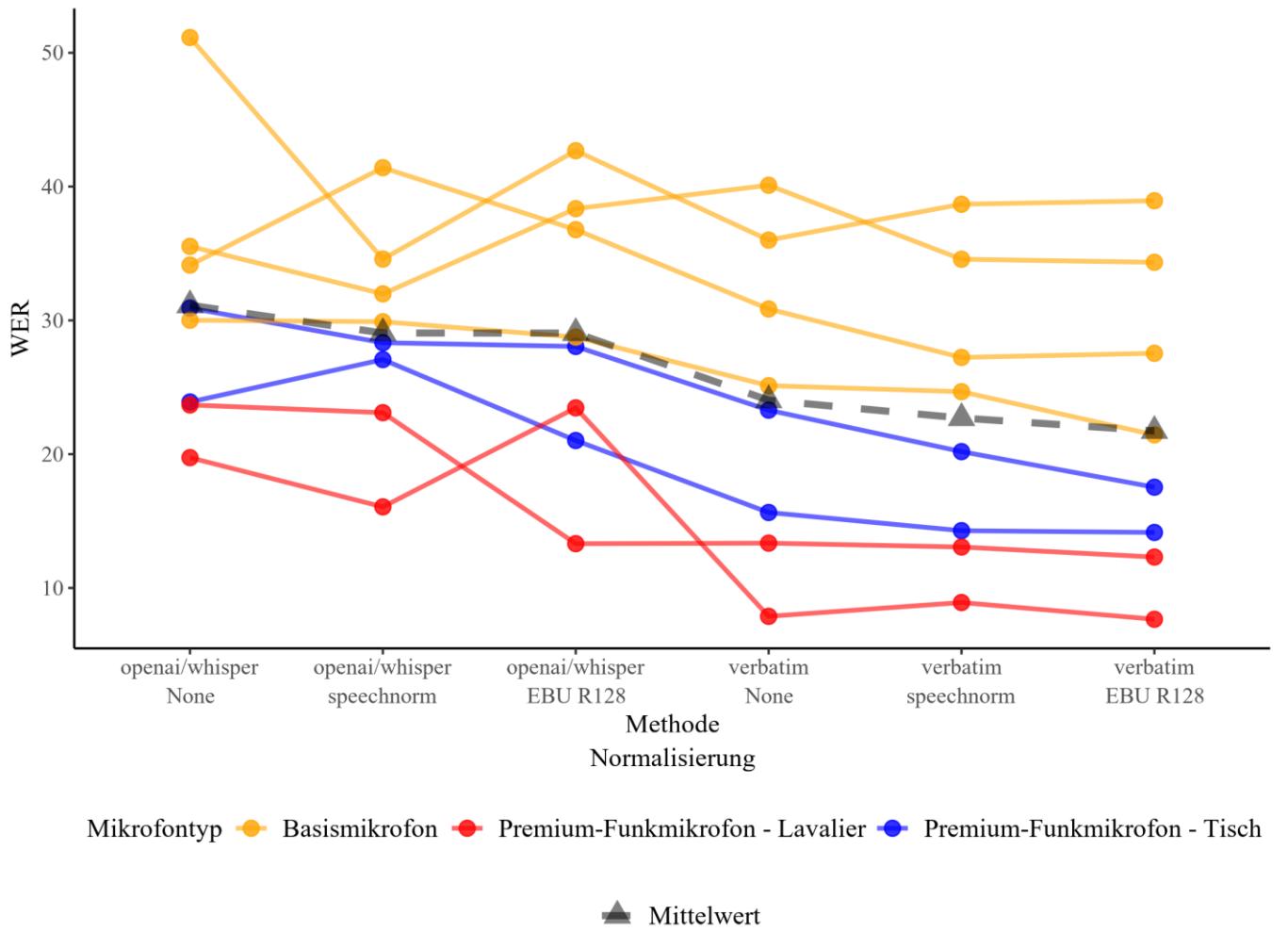
Anmerkungen: *WER = Word Error Rate*; eine niedrigere *WER* weist auf eine höhere Transkriptionsgenauigkeit hin. Sprecher\*innen-Zuordnung: *verbatim* = *pyannote*, *openai/whisper* = keine.

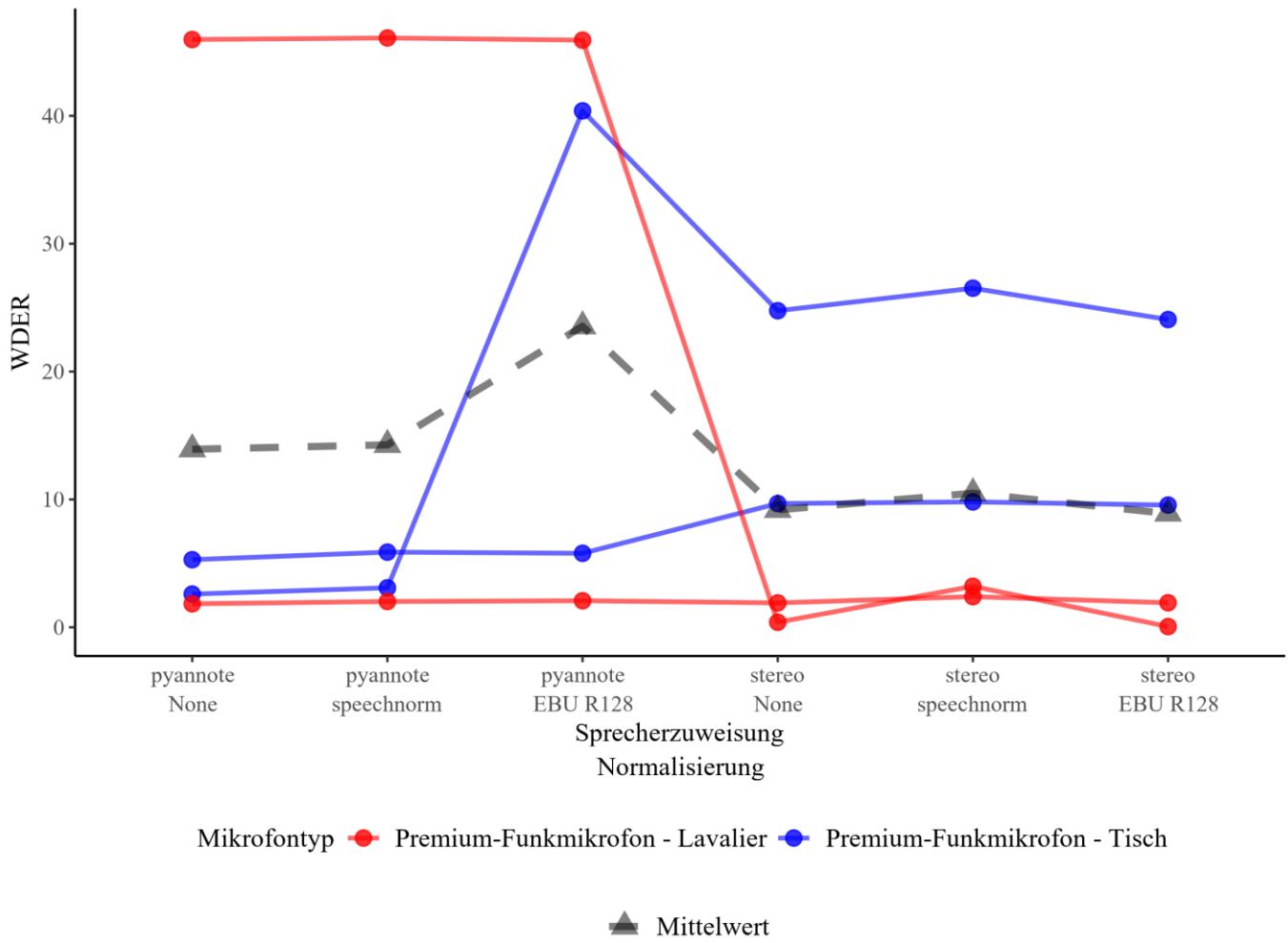
### Abbildung 3. Vergleich der Word Diarization Error Rate (WDER) bei vier Sprecher\*innen-Paaren.

Untersucht werden eine etablierte Methode (*pyannote*) und eine Stereo-Methode unter unterschiedlichen Normalisierungen, Mikrofontypen und Mikrofon-Setups (ohne Basismikrofon).

Anmerkungen: *WDER = Word Diarization Error Rate*; eine niedrigere *WDER* bedeutet eine zuverlässigere Sprecher\*innen-Zuordnung. Für Basismikrofone liegen keine Stereospuren vor, daher entfällt hier der Vergleich.

Prozess	Herausforderung	Lösungsansätze und Methoden in <i>verbatim</i>																		
Tonspuren-Aufbereitung	<ul style="list-style-type: none"> <li>inkonsistente Lautstärke (z.B. durch zu großen Abstand zum Mikrofon, lautes Räuspern etc.)</li> <li>Hintergrundgeräusche (Musik, Lärm etc.)</li> </ul>																			
Transkription	<ul style="list-style-type: none"> <li>fehlerhafte Transkription</li> <li>Konfabulationen (insb. bei Audiosegmenten, die keine Sprache enthalten)</li> <li>Bisher nicht auf (wirtschaftlicher) lokaler Hardware anwendbar</li> </ul>	<p><b>Transkription (whisper-v3-large)</b></p> <p>Stimmaktivitätsanalyse (<i>Silero VAD</i>)</p> <p>Stille wird nicht transkribiert → Reduzierung Konfabulationen</p> <table border="1"> <tr> <td>Hallo, wie geht</td> <td>1a</td> <td>unconfirmed</td> </tr> <tr> <td>Hallo, wie geht</td> <td>es Ihnen?</td> <td>1b confirmed</td> </tr> <tr> <td>Hallo, wie geht</td> <td>es Ihnen?</td> <td>Mir geht 1c</td> </tr> <tr> <td>acknowledged</td> <td></td> <td></td> </tr> <tr> <td>2a</td> <td>Mir geht</td> <td>es gut.</td> </tr> <tr> <td></td> <td></td> <td>usw.</td> </tr> </table>	Hallo, wie geht	1a	unconfirmed	Hallo, wie geht	es Ihnen?	1b confirmed	Hallo, wie geht	es Ihnen?	Mir geht 1c	acknowledged			2a	Mir geht	es gut.			usw.
Hallo, wie geht	1a	unconfirmed																		
Hallo, wie geht	es Ihnen?	1b confirmed																		
Hallo, wie geht	es Ihnen?	Mir geht 1c																		
acknowledged																				
2a	Mir geht	es gut.																		
		usw.																		
Sprecher*innenzuweisung	<p>Deeplearningansatz:</p> <ul style="list-style-type: none"> <li>Bei ähnlichen Stimmprofilen Fehlerhafte Zuweisung (teilweise nicht besser als Zufallszuweisung)</li> <li>Hohe Rechenleistung</li> <li>Mangelnde Nachvollziehbarkeit der Zuordnungsqualität</li> </ul> <p>Lautstärkenanalyse:</p> <ul style="list-style-type: none"> <li>Bedarf Stereo-Kanäle</li> </ul>																			
Evaluation & weiterführende Analysen	<ul style="list-style-type: none"> <li>Abgleich der Transkripte mit Groundtruth</li> <li>Erstellung sinnhafter Aggregationen sowie Wissensextraktion</li> </ul>																			





**Tabelle 1:** Mittlere Prozentuale Word Error Rate (WER) und Word Diarization Error Rate (WDER) im Vergleich

Transkription	Zuordnung	Mikrofon	WER			WDER		
			<i>m</i>	<i>min</i>	<i>max</i>	<i>m</i>	<i>min</i>	<i>max</i>
openai/ whisper	None	BM	37,70	30,01	51,14	NA <sup>*1</sup>	NA <sup>*1</sup>	NA <sup>*1</sup>
		PFM-L	21,71	19,74	23,68	NA <sup>*1</sup>	NA <sup>*1</sup>	NA <sup>*1</sup>
		PFM-T	27,41	23,90	30,92	NA <sup>*1</sup>	NA <sup>*1</sup>	NA <sup>*1</sup>
	<i>pyannote</i>	BM	33,02	25,12	40,10	24,48	2,68	44,35
		PFM-L	10,62	7,88	13,35	23,91	1,84	45,97
		PFM-T	19,46	15,64	23,29	3,95	2,60	5,29
verbatim	stereo	BM	NA <sup>*2</sup>					
	stereo	PFM-L	9,94	7,90	11,98	1,16	0,40	1,91
	stereo	PFM-T	22,00	17,36	26,64	17,22	9,6	24,76

Anmerkung: WER = Word Error Rate, WDER = Word Diarization Error Rate. BM = Basismikrofon, PFM-L = Premium-Funkmikrofon (Lavalier), PFM-T = Premium-Funkmikrofon (Tisch). Keine Normalisierung.

<sup>\*1</sup>Keine Sprecherzuweisung vorgenommen. <sup>\*2</sup>Keine Stereospuren vorhanden.

**Tabelle 2:** Zusammenfassungen und Verdachtsdiagnose von concisum mit Qwen3:8b als LLM: Ausgewählte Befunde aus Anamnese Gesprächen illustrieren jeweils die gepromptete Zusammenfassung, ein Diagnosevorschlag und die dafür angeforderte Begründung. Die Diagnosesicherheit wurde ebenfalls vom Modell eingeschätzt. Die rechte Spalte hebt Beobachtungen typischer Fehler hervor.

Gesprächszusammenfassung und Diagnosevorschlag	Kritische Würdigung der Ergebnisse
<b>Vignette / Zieldiagnose: Alkoholabhängigkeit (F10.2 )</b> <p><b>Therapiesitzung Zusammenfassung</b></p> <p>Die Klient*in berichtet über ein seit 18 Jahren bestehendes Alkoholkonsummuster, das sich in gesellschaftlichen Kontexten sowie nach Stresssituationen (z.B. Streit mit dem Chef) verstärkt. Die Therapeut*in unterstreicht die Bedeutung der Reflexion und der Erkundung von Hintergründen, die den Konsum beeinflussen. Es wird ein Verbindungsmuster zwischen stressinduzierten Emotionen, dem Trinken als Entspannungsmittel und dem damit verbundenen Verhalten (z. B. Streit) identifiziert. Die Klient*in erkennt zunehmend die negativen Auswirkungen des Konsums, einschließlich der Verluste des Führerscheins und der isolierenden Wirkung. Gleichzeitig wird die familiäre Unterstützung durch die Schwester hervorgehoben, die jedoch auch zu Spannungen führt. Die Klient*in beschreibt tiefe Trauer um verstorbene Eltern und eine Furcht vor sozialer Abwertung. Die Therapeut*in validiert die Emotionen und schlägt eine intensivere Bearbeitung der Trauerthemen sowie eine Reflektionsübung zur inneren Prozesse bei Alkoholkonsum vor. Der Fokus liegt auf der Stärkung der Eigenverantwortung und der Balance zwischen Unterstützung und Selbstbeherrschung.</p>	<p>Die Zusammenfassung deckt die wesentlichen Inhalte des Gesprächs ab. (Anmerkung: Das Geschlecht der Sprecher*innen ist in der Transkription der Audiodateien nicht mehr erkennbar.)</p>
<p><b>Diagnosevorschlag</b></p> <p><b>ICD-10 Diagnose</b></p> <p>F10.1 Alkoholabhängigkeit (Alkoholabhängigkeit ohne psychotische Symptome) und F20.0 Schizoides Persönlichkeitsstörung</p> <p><b>Begründung</b></p> <p>Die Symptome der Alkoholabhängigkeit (F10.1) sind direkt nachweisbar durch die kritische Beurteilung der Alkohollaufbahn und die Unfähigkeit, das Verhalten einzuschätzen. Die soziale Isolation und fehlende Selbstreflexion decken zentralen Aspekte der Schizoiden Persönlichkeitsstörung (F20.0) ab, da sie auf ein fehlendes Bedürfnis nach sozialen Beziehungen und eine Unfähigkeit zur Selbstreflexion hinweisen. Die morgendliche Unruhe könnte sekundär auf die Alkoholabhängigkeit zurückgehen (z.B.</p>	<p><i>F10.1 ist falsch, da es einem schädlichen Gebrauch entspricht, (obwohl die Abhängigkeit im Text erkannt wurde). Richtig wäre F10.2 Abhängigkeitssyndrom.</i></p> <p><i>Die Verschlüsselung „Schizoide Persönlichkeitsstörung“ mit F20.0 (paranoide Schizophrenie) ist falsch, korrekt wäre der Schlüssel F60.1, wodurch die diagnostische Trennschärfe verloren geht. Die angeführten Argumente benennen zwar Hinweise, belegen jedoch nicht die vollständigen Persönlichkeitskriterien, sodass eine Persönlichkeitsstörung</i></p>

Entzugssymptome) oder auf eine begleitende Angststörung (F41.1) hindeuten, doch die fehlende kritische Validierung solcher Symptome macht eine zusätzliche Diagnose unsicher. Die Kombination aus F10.1 und F20.0 ist evidenzbasiert, da beide Störungen häufig zusammen auftreten.

**Diagnosesicherheit: 0.85**

unzureichend begründet bleibt.  
Eine behauptete Komorbidität wird ohne differenzialdiagnostische Prüfung und ohne Angabe des zeitlichen Verlaufs aufgestellt. Insgesamt besteht zwar inhaltliche Nähe zur realen/ intendierten Alkoholabhängigkeit, doch die falsche Kodierung und lückenhafte Begründung machen den Vorschlag qualitativ eher mangelhaft.  
Bemerkenswert ist, dass das LLM keine perfekte Sicherheit angibt, also offenbar „erkennt“, dass die Aussagen Lücken aufweisen.

### Identifizierte Symptome

#### Verleugnung (Denial)

**Beschreibung:** Fehlendes Bewusstsein oder Abspruch der eigenen Problemstellung, auch im Kontext von körperlichen oder psychischen Beeinträchtigungen.

**Belege:** [Therapeut]: Ich trinke halt ganz gerne Alkohol und vielleicht kommt es dahin. [Patient]: Nein. [Patient]: Alles nur im geregelten Rahmen.

#### Fehlende Selbstreflexion

**Beschreibung:** Unbewusste oder bewusste Abwehr der eigene Verhaltensmuster und deren Konsequenzen.

**Belege:** [Therapeut]: Ich habe kein Problem. [Therapeut]: Ich trinke schon jeden Tag. [Patient]: Okay, also schon eine ganze Weile dann.

#### Soziale Isolation

**Beschreibung:** Fehlen von sozialen Beziehungen und Unterstützungssystemen, was auf eine Einsamkeit und fehlende soziale Verbindungen hinweist.

**Belege:** Der Therapeut erwähnt, dass seine Frau vor Jahren weggelaufen ist und er seitdem niemanden mehr hat, wodurch sein soziales Umfeld nur aus der Schwester besteht.

#### Alkoholabhängigkeit

**Beschreibung:** Kritik an der Alkohollaufbahn durch externe Quellen und die Unfähigkeit, das eigene Verhalten bewusst einzuschätzen.

**Belege:** Der Therapeut bekommt von seinem Chef und seiner Schwester Rückmeldungen über seinen Alkoholkonsum, obwohl er selbst die Auswirkungen leugnet und denkt, er würde nicht beeinträchtigt.

#### Morgendliche Unruhe

**Beschreibung:** Physiologische Symptome wie morgendliche Unruhe, die auf Angststörungen oder Depressionen hinweisen können.

**Belege:** Der Therapeut sagt: 'morgens ein bisschen

Die Aufstellung greift einige relevante Beobachtungen auf, vermischt jedoch Phänomen-, Mechanismus- und Diagnoseebene. „Verleugnung“ und „fehlende Selbstreflexion“ bezeichnen Einsichtsdefizite, gelten im engeren Sinn nicht als eigenständige Symptome innerhalb der psychopathologischen Nomenklatur. „Alkoholabhängig“ ist wiederum eine Störung (ICD-10 F10.2) und sollte daher nicht als Symptomrubrik geführt werden; stattdessen fehlten in der Liste zentrale Kernsymptome wie Toleranzentwicklung, Craving oder Entzug. „Soziale Isolation“ kann als psychosoziale Folge des Konsums beschrieben werden, ist aber unspezifisch, und die „morgendliche Unruhe“ bleibt ohne Zuordnung zu Entzug oder Angstzustand zu vage.

In Summe ist der Vorschlag teilweise zutreffend, leidet jedoch unter unscharfer Terminologie und lässt wesentliche diagnostisch relevante Merkmale der Alkoholabhängigkeit unberücksichtigt.

Darüber hinaus verwechselt das Modell bei den Belegten Aussagen von der/dem Therapeut\*in mit denen von der/ dem Patient\*in (z.Bsp.: „[Therapeut]: Ich trinke schon jeden Tag.“)

unruhig', was eine körperliche oder emotionale Unruhe im frühen Morgen beschreibt.

### Vignette / Zieldiagnose: Somatisierungsstörung (F45.0)

#### Therapiesitzung Zusammenfassung

Die Klient\*in berichtet über chronische körperliche Beschwerden wie Herzschmerzen, Atemnot und Gelenkschmerzen, die seit Jahren bestehen und das Alltagsleben stark beeinträchtigen. Die Symptome sind nicht auf eine klare organische Ursache zurückzuführen, sondern werden im Kontext von übermäßiger Verantwortung und chronischem Stress thematisiert. Die Klient\*in führt ein vielschichtiges Stressprofil mit: berufliche Pflichten, Haushaltsarbeiten, Pflege der Schwiegermutter sowie ein schuldbewusstes Gefühl im Umgang mit Verpflichtungen. Diese Belastung wird auf verinnerlichte Rollen aus der Kindheit zurückgeführt, wo die Klient\*in als Älteste die Familie im Stich gelassen fühlte. Die Therapeut\*in betont die Verbindung zwischen psychischem Stress und körperlichen Reaktionen und schlägt Mindfulness-Übungen sowie Atemtechniken als Bewältigungsstrategien vor. Die Klient\*in bestätigt die Notwendigkeit weiterer Reflexion und vereinbart eine fortlaufende Therapie. Der Fokus liegt auf der Erkundung von Stressoren, der Aufdeckung von Muster der Überforderung und der Entwicklung von Unterstützungsmechanismen.

*Die Zusammenfassung deckt die wesentlichen Inhalte des Gesprächs ab.*

#### Diagnosevorschlag

##### ICD-10 Diagnose

F48.1 Somatic Symptom Disorder (Schweregrad: mittel) mit begleitenden Störungen der Angst (F41.1) und Depression (F32.0)

##### Begründung

Die Diagnose F48.1 (Somatic Symptom Disorder) ist gerechtfertigt, da die Beschwerden mehrfach auftreten, keine medizinische Ursache haben und zu erheblicher Belastung und Funktionsbeeinträchtigung führen (ICD-10-Kriterium 1-3). Die Angststörung (F41.1) entspricht den Kriterien für übermäßige Sorge um körperliche Beschwerden und Leistungsbeeinträchtigung durch Angst vor gesundheitlichen Folgen. Die depressive Stimmung mit Schuldgefühlen und sozialer Belastung weist auf eine Major Depressive Episode (F32.0) hin, da die Symptome die klinischen Kriterien für eine Major Depression erfüllen (mindestens zwei der fünf Symptome, einschließlich Schuldgefühle). Gütegefühl und emotionale Verdrängung sind zentral für die Depressionsdiagnose. Die Somatisierung (F48.0) wurde nicht als Hauptdiagnose gewählt, da die Symptome im Vordergrund der somatisch orientierten Störung stehen.

**Diagnosesicherheit:** 0.95

*ICD-10-Kodierung falsch: F48.1 bezeichnet das Depersonalisations-/Derealisationssyndrom Richtig wäre für die Die Somatisierungsstörung F45.0 (bzw. F45.1 für die undifferenzierte Form). Die Vermischung des DSM-5-Begriffs „Somatic Symptom Disorder“ mit einem ICD-10-Code führt somit zu einer Fehlklassifikation und reduziert die diagnostische Trennschärfe. Die zusätzlich genannten F41.1 und F32.0 könnten zwar parallel auftreten werden in der Begründung jedoch nicht überzeugend untermauert. Weiterhin ist „Somatisierung (F48.0)“ falsch, da F48.0 Neurasthenie kodiert. Insgesamt liegt der Vorschlag thematisch nah am tatsächlichen Störungsbild, verfehlt aber durch Kodier- und Terminologiefehler die notwendige Präzision.*

	<i>Das LLM ist hier offenbar ein wenig zu selbstsicher.</i>
<b>Identifizierte Symptome</b>	
<b>Somatic Symptom Disorder</b>	
<b>Beschreibung:</b> Mehrfach auftretende körperliche Beschwerden ohne medizinische Ursache, die zu erheblicher Belastung und Funktionsbeeinträchtigung führen	<i>Die Liste spiegelt zwar den zentralen Befund zahlreicher, medizinisch ungeklärter Beschwerden gut wider und steht damit inhaltlich nahe an einer Somatisierungsstörung, vermischt jedoch Störungs- mit Symptomebene. Falsch ist „Somatic Symptom Disorder“ als Symptom statt DSM-5 Diagnose. Ebenso falsch „Angststörung“ ohne eigenständigen Diagnosecode (ICD-10 F41); hier wären präzisere Symptombegriffe wie Krankheitsangst oder anhaltende Beschäftigung über die körperliche Gesundheit erforderlich. Der Terminus „Gütegefühl“ ist eine fehlerhafte Übersetzung von „guilt“ und sollte korrekt als Schuld- oder Versagensgefühle bezeichnet werden, zudem überschneidet er sich weitgehend mit der Rubrik „Depressive Stimmung“. Weiterhin erscheint „Körperliche Beschwerden mit psychischer Ursache“ lediglich als Umschreibung derselben Sachverhalte und trägt wenig zusätzliche Trennschärfe bei. Insgesamt wird die Kernproblematik getroffen, die Benennung bleibt jedoch uneinheitlich und teils redundant sowie terminologisch nicht durchweg fachgerecht.</i>
<b>Belege:</b> Der Patient beschreibt Herzschmerzen, Herzrasen, Atemnot, Gelenkschmerzen, Bauchschmerzen und häufiges Wasserlassen. Die Symptome wurden in acht Krankenhäusern untersucht, ohne eine Erklärung zu finden. Die medizinische Ursache bleibt ungeklärt, was typisch für somatoforme Störungen ist.	
<b>Angststörung</b>	
<b>Beschreibung:</b> Übermäßige Sorge um körperliche Beschwerden und Leistungsbeeinträchtigung durch Angst vor gesundheitlichen Folgen	
<b>Belege:</b> Der Patient äußert sich besorgt über die Unfähigkeit, aus dem Haus zu gehen, und beschreibt das Gefühl, 'eigentlich nichts' zu haben. Die Angst vor körperlichen Symptomen und der Druck, trotzdem arbeiten zu müssen, zeigen eine Angststörungspotenzial.	
<b>Depressive Stimmung</b>	
<b>Beschreibung:</b> Schuldgefühle und Belastung durch soziale Verpflichtungen, die mit emotionaler Verdrängung einhergehen	
<b>Belege:</b> Der Patient erwähnt Schuldgefühle, wenn er seiner Schwiegermutter nicht helfen kann, und beschreibt das Gefühl, 'schlimm' zu sein, wenn er ausfällt. Dies deutet auf depressive Symptome wie Schuldgefühle und funktionale Beeinträchtigung hin.	
<b>Gütegefühl</b>	
<b>Beschreibung:</b> Übermäßiges Schuldgefühl oder Schuldbezogene Empfindungen in Situationen, in denen man sich nicht um etwas kümmern kann (z. B. Krankheit, Pflegepflichten).	
<b>Belege:</b> Der Therapeut beschreibt, dass er sich schlechtes Gewissen macht, wenn er krank ist, weil Kollegen einspringen müssen, und dass er sich schuldig fühlt, wenn seine Familie allein bleibt.	
<b>Körperliche Beschwerden mit psychischer Ursache (Somatisierung)</b>	
<b>Beschreibung:</b> Körperliche Symptome, die nicht direkt medizinisch nachvollziehbar sind, aber psychische Belastungen wie Stress oder Überforderung reflektieren.	
<b>Belege:</b> Der Therapeut erwähnt, dass er körperliche Schmerzen verspürt, die er nicht einbildet, aber nicht nachvollziehen kann, was auf eine verminderte Körperwahrnehmung oder Stresssymptome hinweist.	

## Vignette / Zieldiagnose: Posttraumatische Belastungsstörung (F43.1)

### Therapiesitzung Zusammenfassung

Die Klient\*in beschreibt schwere psychische Belastungen nach einem Autounfall mit der Nichte, der zu Schlafstörungen, Angstzuständen, Panikattacken und einer Vermeidung von Autofahren führte. Die Verletzung der Nichte und die Erinnerung an die Mutter, die an einem Unfall starb, verstärkten Schuldgefühle und Trauer. Die Klient\*in leidet unter sozialer Isolation, Konzentrationsschwächen und einem veränderten Alltag, der die berufliche und private Lebensqualität beeinträchtigt. Die Therapeut\*in betont die tiefgreifende Belastung durch die Vorfälle sowie die Wiedererleben von Traumata. Die Klient\*in signalisiert Bereitschaft zur Therapie, um Schuldgefühle zu verarbeiten, die soziale Kontakte wiederzuerlangen und den Lebensrhythmus zu stabilisieren. Die Therapie wird als kognitive Verhaltenspsychotherapie eingesetzt, um Emotionen, Gedanken und Verhalten zu bearbeiten. Die Klient\*in betont die Bedeutung von Verständnis und Sicherheit in der Therapie, unterstützt durch Supervision. Die Arbeitsbelastung hat sich erhöht, da die Klient\*in nicht mehr autofahren kann, was zu längeren Arbeitszeiten führt. Die Therapeut\*in versichert, den Prozess mit Sicherheit und Qualität zu gestalten.

Die Zusammenfassung deckt die wesentlichen Inhalte des Gesprächs ab. Die Zusammenfassungsstruktur orientiert sich dabei an den inhaltlichen Themen entlang der Sitzung und erfassst ebenfalls Gesprächsredundanzen.

### Diagnosevorschlag

#### ICD-10 Diagnose

F43.1 Posttraumatische Störung (PTSD) mit Insomnie (F51.0) und irritablem Stimmung, ggf. begleitende depressive Episode (F32.1) oder Delusionszüge (F22.9).

#### Begründung

1. Posttraumatische Störung (F43.1): Die Beschreibung der emotionalen Überlastung durch ein traumatisches Ereignis (Unfall) und die damit verbundenen psychischen Belastungen (angsthaftes Verhalten vor Therapiesitzungen, Schlafstörungen, Reizbarkeit) decken die Kriterien für PTSD ab. Der Trauma-Auslöser (Unfall) ist explizit genannt, und die Symptome umfassen intrusivere Erinnerungen, Überwachungsangst (Schlafstörung) und veränderte Reaktionsmuster (Irritabilität).

2. Insomnie (F51.0): Die Unfähigkeit, sich zu entspannen und zu schlafen, mit emotionaler Überlastung entspricht den Kriterien für eine chronische Schlafstörung. Die Schlafprobleme sind als zentraler Begleitsymptom der PTSD dokumentiert.

3. Irritabilität: Die Zusammenhangslose Reizbarkeit und Frustration („voll generiert“) sind typische symptomatische Begleiterscheinungen von PTSD und können auch bei depressiven Episoden auftreten. Da die

Posttraumatischen Belastungsstörung F43.1 trifft die korrekte/intendierte Hauptdiagnose. Falsch ist die zusätzliche Kodierung einer Insomnie (F51.0), da Ein- und Durchschlafstörungen als Leitsymptom bereits in F43.1 enthalten sind und so eine Doppelkodierung auf Symptom- versus Störungsebene entsteht. Die ggf. mittelgradige depressive Episode (F32.1) wird ohne Prüfung aller Haupt- und Zusatzsymptome sowie des Zwei-Wochen-Kriteriums nur behauptet und bleibt deshalb unzureichend begründet. Die Angabe von Delusionszügen (Übersetzungsfehler; korrekt wäre anhaltende wahnhafte Störung, nicht näher bezeichnet) mit F22.9 überzeugt nicht, weil zeitliche Dauer und systematischer Wahngehalt fehlen und damit eine Fehleinschätzung droht. Insgesamt ist der Vorschlag dank der korrekten Kernkodierung

emotionale Überlastung im Kontext des Traumas steht, wird dies als Teil der PTSD betrachtet.	fachlich solide, verliert aber durch Übercodierung und schwach belegte Zusatzdiagnosen an Präzision.
4. Delusionszüge (F22.9): Die Selbstbezogene Wahnidee („Ich spinne, weil ich kann nicht mehr richtig schlafen“) ist ein indirekter Hinweis auf Delusionszüge. Allerdings fehlen die zentralen Kriterien für eine Delusionsstörung (z.B. fehlende Halluzinationen, kognitive Störungen), sodass dies als unklare Begleiterscheinung des PTSD oder einer depressiven Episode mit psychotischen Merkmalen (F32.1) angesehen wird. Die Sicherheit der Diagnose ist aufgrund der indirekten Formulierung der Wahnvorstellung begrenzt.	<i>Die Selbstsicherheit des LLM scheint gerechtfertigt.</i>

**Diagnosesicherheit:** 0.85

#### Identifizierte Symptome

##### Angststörung

**Beschreibung:** Erhöhte Aufregung und Angst vor der Therapiesitzung, die sich in der Aussage 'Ich bin nur ganz arg aufgeregt' widerspiegelt.

**Belege:** Ich bin nur ganz arg aufgeregt.

##### Schlafstörung

**Beschreibung:** Unfähigkeit, sich richtig zu entspannen und zu schlafen, verbunden mit emotionaler Überlastung.

**Belege:** Ich kann nicht mehr richtig schlafen.

##### Irritabilität

**Beschreibung:** Zusammenhangslose Reizbarkeit und Frustration, die sich in der Aussage 'Ich bin immer gereizt, ich bin voll genervt' widerspiegelt.

**Belege:** Ich bin immer gereizt, ich bin voll genervt, es sind so viele Sachen.

##### Emotionale Überlastung

**Beschreibung:** Überforderung durch traumatische Ereignisse (Unfall) und die damit verbundenen psychischen Belastungen.

**Belege:** Ich höre jetzt schon, dass sie das sehr belastet und mitnimmt. Es ist einfach scheiße.

##### Wahnvorstellung

**Beschreibung:** Selbstbezogene Wahnideen, die in der Aussage 'Ich spinne, weil ich kann nicht mehr richtig schlafen' implizit formuliert werden.

**Belege:** Ich spinne, weil ich kann nicht mehr richtig schlafen.

*Die Liste enthält zwar einige für eine Posttraumatische Belastungsstörung relevante Phänomene (z. B. Schlafstörung und Irritabilität), verwendet jedoch uneinheitliche Ebenen: „Angststörung“ ist eine Diagnose und kein Symptom, während „emotionale Überlastung“ als unspezifischer Befund ohne klaren psychopathologischen Terminus bleibt. Die Interpretation der Aussage „Ich spinne ...“ als „Wahnvorstellung“ ist falsch, da hierfür fixe, realitätsferne Überzeugungen mit fehlender Korrekturfähigkeit nachzuweisen wären. Zudem fehlen zentrale PTBS-Merkmale wie intrusive Wiedererleben-Phänomene, Vermeidung oder Hypervigilanz, sodass das Bild unvollständig bleibt. Insgesamt zeigt der Vorschlag Ansätze richtiger Beobachtung, leidet aber unter terminologischen Ungenauigkeiten und wesentlichen Auslassungen, wodurch die diagnostische Trennschärfe eingeschränkt ist.*

**Anmerkung.** Die Tabelle zeigt drei automatisiert transkribierte und zusammengefasste Therapiegespräche sowie ein Diagnosevorschlag respektive Symptomextraktion mittels eines Multi-Shot-Agent-Systems unter der Verwendung von Qwen3:8b. Diagnosesicherheit gibt die vom Modell geschätzte Wahrscheinlichkeit der korrekten Diagnose (0 = sehr unsicher, 1 = sehr sicher) an.